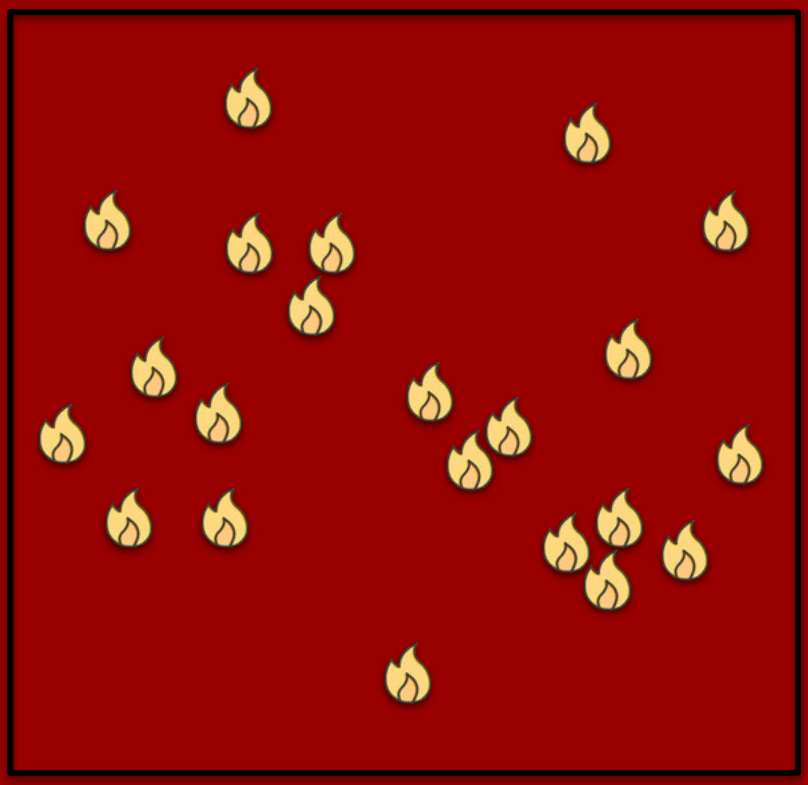
Statistical and Machine Learning Contributions to Spatial and Spatio-temporal Point Process Modelling, with an Application to Dutch Fire Risk Prediction



Changqing Lu

STATISTICAL AND MACHINE LEARNING CONTRIBUTIONS TO SPATIAL AND SPATIO-TEMPORAL POINT PROCESS MODELLING, WITH AN APPLICATION TO DUTCH FIRE RISK PREDICTION

Changqing Lu

The research leading to this dissertation has received funding from the Dutch Research Council for the project 'Data driven risk management for fire services' (18004). The data collection was conducted in collaboration with the Twente Fire Brigade.

STATISTICAL AND MACHINE LEARNING CONTRIBUTIONS TO SPATIAL AND SPATIO-TEMPORAL POINT PROCESS MODELLING, WITH AN APPLICATION TO DUTCH FIRE RISK PREDICTION

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. ir. A. Veldkamp,
on account of the decision of the Doctorate Board,
to be publicly defended
on Thursday 27th February 2025 at 12.45 hours

by

Changqing Lu

born on the 2nd of November, 1997
in Taizhou, Jiangsu, China

This dissertation has been approved by:

Promotors
prof. dr. M. N. M. van Lieshout
prof. dr. R. J. Boucherie

Co-promotor
dr. ir. M. de Graaf

Cover design: Changqing Lu
Printed by: Ipskamp Printing
Lay-out: Changqing Lu
ISBN(print): 978-90-365-6497-7
ISBN(digital): 978-90-365-6498-4
DOI: https://doi.org/10.3990/1.9789036564984

© 2025 Changqing Lu, Enschede, The Netherlands.

All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author. Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd, in enige vorm of op enige wijze, zonder voorafgaande schriftelijke toestemming van de auteur.

Graduation Committee:

Chair / secretary:	prof. dr. ir. B. R. H. M. Haverkort
Promotors:	prof. dr. M. N. M. van Lieshout CWI & University of Twente
	prof. dr. R. J. Boucherie University of Twente
Co-promotor:	dr. ir. M. de Graaf Thales B.V. & CWI
Committee Members:	prof. dr. A. J. Schmidt-Hieber University of Twente
	prof. dr. A. Stein University of Twente
	dr. R. Bardenet CNRS, CRIStAL & University of Lille, France
	prof. dr. G. Jongbloed Delft University of Technology
	prof. dr. R. P. Waagepetersen Aalborg University, Denmark
	dr. ir. R. A. C. de Wit Twente Fire Brigade

Acknowledgements

The completion of my PhD has been an incredible journey, made possible by the guidance, support and encouragement of many remarkable individuals and groups. I would like to take this opportunity to express my heartfelt gratitude to everyone who contributed to this endeavor.

First, I am deeply grateful to my (co-)promotors. Marie-Colette, thank you for introducing me to the world of statistical academia, for inspiring a deep passion in me on the research I have been conducting, for your patience and efforts in answering all my, sometimes even naive, questions, for your concerns not only on my research but also on my daily life, and for always standing behind to support me throughout this journey. Richard, I sincerely appreciate your kindness, and the positive, encouraging working environment you have fostered, which greatly contributed to my growth as a researcher. Maurits, your thoughtful comments and active involvement in our research projects are valuable. Even after you leaving the University of Twente midway through my PhD, you remained engaged, and your constructive input and steady oversight of my progress were instrumental to my development.

Second, I am highly grateful to my oversea collaborators for their precious contributions. Yongtao, thank you for insightful ideas and guidance on significant advancements in the domain of spatial statistics and for graciously hosting me during my research visit to the Chinese University of Hong Kong. Ganggang, I am particularly thankful for your detailed, inspiring suggestions on our work, as well as your tireless and meticulous revisions. I will always remember the moments we improved the papers together, despite the challenges of time differences across Europe, United States and China.

Third, I would like to extend my sincere gratitude to the committee members of my graduation for dedicating your time to reviewing my thesis and providing useful feedback that helps improve my work. Furthermore, I would also like to thank the user committee of my PhD project and the business intelligence department of Twente Fire Brigade, especially Ron, Paul, Niels and Emiel. The development of our methodologies and theories on spatial and spatio-temporal modelling, proposed in this thesis, could not be achieved without your motivations, data collection efforts and constructive comments from the practical perspective.

Fourth, I want to extend my gratitude to colleagues in the Mathematics of Operations Research group at the University of Twente for your warm conversations. To my fellow PhD candidates in spatial statistics and stochastics, Zhuldyzay,

Robin and Maike, thank you for your companionship in the reading group, exchanging ideas and sharing each other's research experiences. I am also grateful to the PhD candidates in the Statistics group for enriching discussions on various statistical problems and for making my research life enjoyable. I want to express special thanks to Clement and Hongwei for brainstorming research questions that potentially inspired part of my thesis. To my officemates, Nhung, Matthew, Anouk, Leon and Dennis, thank you for joyful conversations and shared humor that made our office vibe pleasant. Another special thank you to Thyra and Marjo, for your efforts in providing administrative support. Finally, I would like to extend my gratitude to friends at the University of Twente, in Enschede, and across the Netherlands, the Dutch, the Chinese and the internationals. Thank you for the wonderful stories we created together, from sports games to social gatherings, which made my journey truly unforgettable.

Last but not least, my deepest thanks go to my family, my father, mother and girlfriend. Thank you for your love, encouragement and support, which have been the most important foundation of my success in this journey.

Contents

CHAPTER 1

Introduction

Spatial and spatio-temporal point process models have been widely applied in various disciplines, such as geology (e.g., Unwin, 1996; Bray & Schoenberg, 2013), epidemiology (e.g., Gatrell et al., 1996; Dong et al., 2023; Schoenberg, 2023), ecology (e.g., Stoyan & Penttinen, 2000; Waagepetersen, 2008) and society (e.g., Xu, Luo & Zha, 2017; D'Angelo et al., 2022; Zhu & Xie, 2022), to model the occurrence of random events, such as earthquakes, infections, species distributions and crimes, in space and/or time. These models are particularly suitable for analyzing point pattern data, where points are observed within some specific window (see, examples in Figure 1.1), and the objective is to understand the underlying mechanisms that govern the spatial and/or temporal distributions of these points.

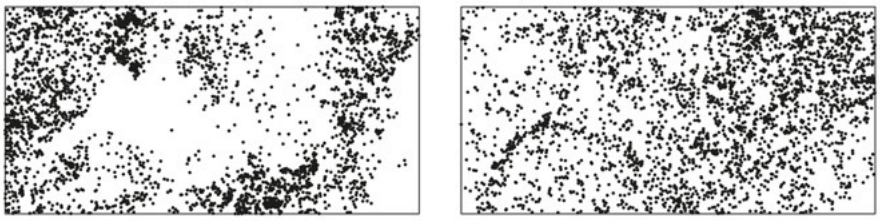


Figure 1.1 Point patterns of locations of two tree species, Beilschmiedia pendula and Capparis frondosa, on Barro Colorado Island in Panama (Condit et al., 2019).

A key advantage of point process models lies in their probability structure, which not only allows for capturing the spatial and temporal spread of individual events but also for explaining how the events are influenced by the surrounding environment and interact with one another. For example, in ecology, the distribution of tree species observed in a region, as shown in Figure 1.1, is usually determined by environmental factors, such as terrain elevation and slope, cf. Figure 1.2. In geology, the occurrence of a major earthquake is typically followed by multiple aftershocks, and the influence of the main shock on the occurrence of an aftershock can be modelled by some triggering function, e.g. exponential

decay $\exp(-|t - t_0|/\beta)$ with t_0 and t denoting the time of the main shock and the aftershock, respectively, and β denoting a scale parameter > 0. In epidemiology, the spread of diseases, like COVID-19, may be captured by incorporating pairwise interactions within a population network, where each vertex represents an individual, and social relationships are modelled as edges which may activate a vertex to become a point of infection depending on actual physical contacts between individuals.

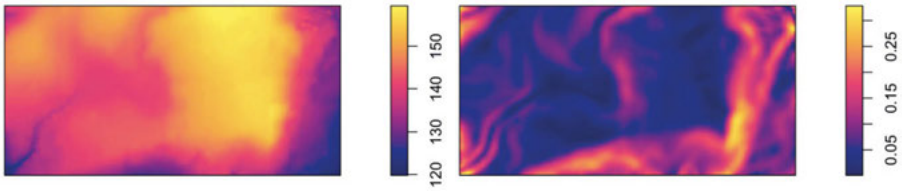


Figure 1.2 Terrain elevation and slope measured in the region of the forestry data displayed in Figure 1.1 (Condit et al., 2019).

By leveraging these increments as introduced above, spatial and spatio-temporal point process models offer both flexibility and interpretability in learning first- and second-order features of point pattern data (cf., Sections 1.1.1 and 1.1.2) and enables us to understand complex spatial and temporal phenomena of occurrences of events in real-world applications.

In particular, this thesis advances methodological and theoretical research in point process modelling, with a focus on its application to fire risk prediction in the Netherlands. Moreover, as an important contribution, it integrates several machine learning techniques to the point process context, addressing certain challenges presented in particular data scenarios that are difficult to solve by traditional statistical approaches.

1.1 Point Process Theory

Generally speaking, a point process is a mathematical model that describes the occurrence of points at random locations within a given space (Daley & Vere-Jones, 2009). More formally, to define a point process X, we shall start with a probability space $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ is the event space, and P is a probability function assigning probabilities to the events in $\mathcal{F}$. We also need a measurable space $(\mathcal{S}, \mathcal{B})$, where $\mathcal{S}$ is a complete, separable metric space and $\mathcal{B}$ is its Borel σ-algebra. Under this framework, X is a mapping from $(\Omega, \mathcal{F}, P)$ into $(\mathcal{S}, \mathcal{B})$ and can be further considered as an integer-valued random measure $N(\cdot)$ on $\mathcal{S}$ that counts the number of points in a subset of $\mathcal{S}$.

Similar to most of the literature, we restrict ourselves to point processes whose realizations are locally finite subsets of $\mathcal{S}$. To define such point configurations, we denote the cardinality of a subset $x \subset \mathcal{S}$ by $n(x)$. For every bounded set $B \subset \mathcal{S}$, x is said to be locally finite if $n(x_B) < \infty$, where x_B is the restriction of

x to B. Denote the family of the locally finite point configurations that consist of all subsets of $\mathcal{S}$ which place finitely many points in every bounded Borel set $B \subset S$ by N_{lf}. The following definition of point processes is used throughout the thesis.

Definition 1.1. *A point process X on $\mathcal{S}$ is a mapping from a probability space $(\Omega, \mathcal{F}, P)$ into N_{lf} such that, for all bounded Borel sets $B \subset S$, the number of points of X that fall in B, denoted by $N(B)$, is a finite random variable.*

For spatial point processes, $\mathcal{S} \subset \mathbb{R}^2$, while in the spatio-temporal context, $\mathcal{S}$ is typically a bounded observation window in space and time, which we denote by $\mathcal{W} \times \mathcal{T} \subset \mathbb{R}^2 \times \mathbb{R}^+$ with $\mathcal{W}$ representing the spatial domain and $\mathcal{T}$ denoting the temporal interval. Consequently, X becomes a random countable subset of $\mathcal{W} \times \mathcal{T}$.

There are various point process models, with the Poisson point process being the most fundamental one, as it assumes no interactions between points. Specifically, it is defined as follows.

Definition 1.2. *A point process X on $\mathcal{S}$ is called a Poisson process if it satisfies the following two conditions: (i) for any bounded Borel set $B \subset \mathcal{S}$, $N(B)$ is Poisson distributed with a mean proportional to the Lebesgue measure of B; (ii) for any disjoint bounded Borel sets $B_1, B_2 \subset \mathcal{S}$, $N(B_1)$ and $N(B_2)$ are independent from each other.*

These properties are known as 'complete spatial randomness' in statistical theory, allowing the Poisson point process to act as a reference process when testing summary statistics for complex model structures.

More advanced point process models introduce dependencies between points, enabling suitable modelling for different types of point pattern data. A well-known example is the log-Gaussian Cox process (Coles & Jones, 1991; Møller, Syversveen & Waagepetersen, 1998), where the occurrence of points within the observation window is driven by a latent Gaussian random field. It introduces spatial and temporal correlations to the model such that nearby locations and times exhibit similar point patterns, whereas distant locations are less dependent. Another common example is the Markov point process (Ripley & Kelly, 1977; Van Lieshout, 2000), in which the occurrence of new points is influenced by the presence of existing events. Depending on the selected interaction function, such models can capture either clustered or repulsive phenomena in data.

1.1.1 First-order Features

The first-order features of a point process are characterized by its intensity function. To introduce it, we need the definition of the first-order moment measure.

Definition 1.3. *The first-order moment measure of a point process X on $\mathcal{S}$ is defined as*

$$\Lambda(B) = \mathbb{E}\left[N(B)\right] = \mathbb{E}\left[\sum_{\boldsymbol{x} \in X} 1(\boldsymbol{x} \in B)\right] \tag{1.1}$$

for any bounded Borel set $B \subset \mathcal{S}$.

Here, $1(\cdot)$ denotes the indicator function and the sum is over all point realizations in X. Suppose that Λ exists as a σ-finite measure that is absolutely continuous with respect to Lebesgue measure with Radon–Nikodym derivative λ. Then, $\lambda : \mathcal{S} \to [0, \infty)$ is called the intensity function of X.

Alternatively, the intensity function can be explained in a more intuitive way (Diggle, 2013). Let $|B|$ denote the area of B and let $\mathrm{d}\boldsymbol{s}$ denote an infinitesimal ball centred at $\boldsymbol{s} \in \mathcal{S}$. The intensity $\lambda(\boldsymbol{s})$ then represents the approximated probability for $\mathrm{d}\boldsymbol{s}$ to contain exactly one point from X. Hence,

$$\lambda(\boldsymbol{s}) = \lim_{|\mathrm{d}\boldsymbol{s}| \to 0} \mathbb{E}\left[\frac{N(\mathrm{d}\boldsymbol{s})}{|\mathrm{d}s|}\right]. \tag{1.2}$$

For Poisson point processes, its probability distribution is completely specified by the intensity function. Under the conditions in Definition 1.2, the likelihood of a Poisson point process X is given by

$$\left[\prod_{\boldsymbol{x} \in X} \lambda(\boldsymbol{x})\right] \exp\left[-\int_{\mathcal{S}} \lambda(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}\right]. \tag{1.3}$$

When $\lambda(\boldsymbol{s})$ is a constant for all $\boldsymbol{s} \in \mathcal{S}$, X is said to be a homogeneous Poisson point process. Moreover, if $\lambda(\boldsymbol{s}) \equiv 1$, it is called the unit-rate Poisson point process.

1.1.2 Second-order Features

The second-order features of a point process refer to statistical properties that describe point dependencies between pairs. More than the intensity function, it provides deeper insights into how points interact with one another, particularly in terms of clustering or regularity. In analogy to Definition 1.3, we shall need the following definition of the second-order factorial moment measure.

Definition 1.4. *The second-order factorial moment measure of a point process X on $\mathcal{S}$ is defined as*

$$\Lambda^{(2)}(B_1 \times B_2) = \mathbb{E}\left[\sum_{\boldsymbol{x} \in X} \sum_{\boldsymbol{y} \in X, \boldsymbol{y} \neq \boldsymbol{x}} 1(\boldsymbol{x} \in B_1, \boldsymbol{y} \in B_2)\right] \tag{1.4}$$

for any bounded Borel sets $B_1, B_2 \subset \mathcal{S}$.

Here, the sum is over all pairs of distinct points in X. Suppose that $\Lambda^{(2)}$ exists as a σ-finite measure that is again absolutely continuous with respect to Lebesgue measure with Radon–Nikodym derivative $\lambda^{(2)}$. Then, $\lambda^{(2)} : \mathcal{S}^2 \to [0, \infty)$ is called the second-order product density function of X.

Similarly, the second-order product density function at locations $\boldsymbol{x}$ and $\boldsymbol{y}$ can be explained as the approximated probability that X places a point in each of the two infinitesimal balls centred at $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$. Hence,

$$\lambda^{(2)}(\boldsymbol{x}, \boldsymbol{y}) = \lim_{|\mathrm{d}\boldsymbol{x}|,|\mathrm{d}\boldsymbol{y}|\to 0} \mathbb{E}\left[\frac{N(\mathrm{d}\boldsymbol{x})N(\mathrm{d}\boldsymbol{y})}{|\mathrm{d}\boldsymbol{x}||\mathrm{d}\boldsymbol{y}|}\right]. \tag{1.5}$$

For Poisson point processes, $\lambda^{(2)}(\boldsymbol{x}, \boldsymbol{y}) = \lambda(\boldsymbol{x})\lambda(\boldsymbol{y})$.

One of the most commonly used summary statistics to assess the second-order feature of a point process is the pair correlation function, which measures how likely two points in X are to occur together compared to what would be expected under the assumption of Poisson point processes, i.e. completely independent. Suppose that both Λ and $\Lambda^{(2)}$ exist, the pair correlation function of a point process X is defined as

$$g(\boldsymbol{x}, \boldsymbol{y}) = \frac{\lambda^{(2)}(\boldsymbol{x}, \boldsymbol{y})}{\lambda(\boldsymbol{x})\lambda(\boldsymbol{y})} \tag{1.6}$$

provided that $\lambda(\boldsymbol{x})\lambda(\boldsymbol{y}) > 0$. Furthermore, if one assumes second-order intensity-reweighted stationarity (Baddeley, Møller & Waagepetersen, 2000), (1.6) can be simplified to $g(\|\boldsymbol{x} - \boldsymbol{y}\|)$ which only depends on the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Another popular statistics for detecting second-order features is the K-function, also known as Ripley's reduced second moment function (Ripley, 2012). It examines the expected number of events within a given distance from any arbitrary point, divided by the intensities. The Ripley's K-function of an inhomogeneous spatial point process X is defined as

$$K(r) = \frac{1}{|B|}\mathbb{E}\left[\sum_{\boldsymbol{x}\in X\cap B}\sum_{\boldsymbol{y}\in X, \boldsymbol{y}\neq\boldsymbol{x}} \frac{1(||\boldsymbol{x} - \boldsymbol{y}|| \leq r)}{\lambda(\boldsymbol{x})\lambda(\boldsymbol{y})}\right]. \tag{1.7}$$

For spatio-temporal point processes, one needs to separate the distance above in space and time. In an inhomogeneous version, it takes the form of

$$K_{\text{inhom}}(r, v) = \frac{1}{|B|}E\left\{\sum_{\boldsymbol{x}\in X\cap B}\sum_{\boldsymbol{y}\in X, \boldsymbol{y}\neq\boldsymbol{x}} \frac{1(||\boldsymbol{w}(\boldsymbol{x}) - \boldsymbol{w}(\boldsymbol{y})|| \leq r, ||t(\boldsymbol{x}) - t(\boldsymbol{y})|| \leq v)}{\lambda(\boldsymbol{x})\lambda(\boldsymbol{y})}\right\}, \tag{1.8}$$

where the spatial distance and temporal interval $r, v \geq 0$, and $\boldsymbol{x}$ and $\boldsymbol{y}$ are distinctive points with $\boldsymbol{w}(\boldsymbol{x}), \boldsymbol{w}(\boldsymbol{y})$ and $t(\boldsymbol{x}), t(\boldsymbol{y})$ denoting their spatial and temporal coordinates.

The pair correlation function and Ripley's K-function play a significant role in point pattern analysis. First, they can be directly used to assess the second-order properties of an observed point pattern. Specifically, a pair correlation function larger than 1 or a spatial Ripley's K-function larger than πr^2 indicates clustering, while the values smaller than one and πr^2 suggest dispersion (note

that, in the spatio-temporal case, πr^2 changes to $2\pi r^2 v$). For ease of understanding, an example of clustered, Poisson and repulsive point patterns is displayed in Figure 1.3. Second, the observed pair correlation function and Ripley's K-function from point pattern data, when equipped with inference techniques like the minimum contrast estimation, can help determine the second-order structure of a proposed point process model (Waagepetersen & Guan, 2009). Last but not least, in statistical testing, Monte Carlo simulations can be conducted based on a fitted model to generate confidence envelopes for the two functions, allowing for validation of model adequacy (Baddeley, Rubak & Turner, 2015).

Figure 1.3 An example of clustered, Poisson and repulsive spatial point patterns in the unit window $\mathcal{S} = [0, 1]^2$.

1.2 Fire Risk Prediction in the Netherlands

The research in this thesis is motivated by the fire risk prediction practices in the Netherlands. In recent decades, the Dutch fire and rescue services have increasingly explored the use of business intelligence to optimize their operations. Key strategy priorities include enhancing public safety, improving financial oversight and strengthening risk management. As a first step, substantial efforts have been devoted to data collection, resulting in an extensive database with detailed information on fire incidents. The ultimate objective is to use these data to generate reliable, actionable predictions for formulating better staffing and equipment arrangements and developing more effective fire prevention measures.

Spatial and spatio-temporal point processes provide a natural framework for modelling fire data (e.g., Møller & Díaz-Avalos, 2010; Serra et al., 2014; Koh et al., 2023), where fire incidents are treated as randomly occurring events, observed across the space and time of interest. The primary goal of fire prediction is thus to estimate the intensity function of the underlying point process, which specifies the fire risk intuitively. Given the critical role of uncertainty quantification in risk management, establishing confidence intervals for fire risk estimates is also important. In a prior study, Wendels (2017) demonstrated that different types of fires exhibit distinct patterns of occurrence, necessitating tailored models to

accurately capture these phenomena. Moreover, the practical complexity of fire risk prediction poses significant challenges to the statistical literature on point process modelling. Through consultations with the firefighters, we identified three common phenomena in fire data. We describe them below in detail and will discuss the associated statistical challenges afterwards.

1.2.1 Covariate-based Phenomenon

The risk of certain fires is dominated by environmental factors, which are referred to as covariates in statistical theory. A typical example are chimney fires, occurring more frequently in residential areas with a larger population and especially in winter. The reasons are straightforward: the population size determines the use of chimneys and the cold weather prompts people's reliance on chimneys for heating. Beyond these basic factors, other variables can also influence the risk of chimney fires. For instance, old chimneys made of brick are more susceptible to fires compared to the new ones constructed from metal. Moreover, modern buildings are equipped with central heating systems and lack chimneys entirely. Temporally, certain weather conditions, e.g. fog, can heighten the risk. For such fire types, analyzing the relationship between fire intensity and related covariates is essential for accurate risk prediction. To comprehensively assess the impact of environmental factors, firefighters have compiled a wide range of putative covariates according to their experience. Specifically, for chimney fires, 27 putative explanatory variables are accessible, cf. Chapter 2. However, not all of them are necessary for effective fire risk modelling. Properly selecting the most important covariates and utilizing them to build an interpretable, accurate fire prediction model is crucial for addressing this type of covariate-based phenomenon.

1.2.2 Random Effect Phenomenon

In addition to the covariate information, some fire risks can also be influenced by random effects, which in statistical theory, account for unexplained spatio-temporal correlations in point pattern data. For example, School (2018) attempted to predict chimney fire risk by incorporating population and temperature as covariates in a homogeneous Poisson point process model. However, validation revealed that the purely covariated-based model underestimated fire occurrences, with residuals exhibiting a clustering phenomenon. To tackle this, a Gaussian random field was introduced as a random effect to model the spatial and temporal residuals. In fact, these correlations could be interpreted as additional covariates, although relevant data are often unavailable. In the case of chimney fires, spatial correlations may arise from shared behaviors among neighbours, such as purchasing high-quality chimneys and regularly cleaning them. Temporally, unexpected storms may also bring short-term dependencies into the fire risk. While including random effects sometimes improves prediction accuracy by capturing the 'hidden' correlations, they are not always necessary, especially when sufficient covariates are available. Instead, model underfitting may result

from insufficiently leveraging the covariate information, underscoring the need to exploit the use of accessible environmental data.

1.2.3 Hierarchically Clustered Phenomenon

Another common phenomenon in fire data is the clustering pattern in a hierarchical structure, which differs from the correlations introduced by random effects as discussed above because it is driven by certain latent cluster centres. A typical example of this phenomenon are car arson fires, which are often committed by a specific group of arsonists around particular activity centres, e.g. frequently in low-income and low-education neighbourhoods. In addition, according to the police, arsonists tend to target specific locations and times, such as outside a bar on a Sunday evening or near a stadium following a football match. For modelling purposes, this clustered phenomenon can be explained in a hierarchical framework of two layers: the first layer represents the latent activity centres of arsonists and the second layer involves the observed fire incidents aggregated from all clusters. Predicting risk for this type of fires requires not only estimating the occurrences of actual incidents across space and time but also identifying the underlying arson clusters based on past fire data. By pinpointing the cluster centres, firefighters, in collaboration with the police, can implement safety measures more effectively in high-risk areas, as arson fires, being criminal acts, pose additional challenges compared to accidental fires.

1.3 Research Problems

This thesis explores statistical problems on spatial and spatio-temporal point process modelling that arise from the fire risk prediction practices discussed above. In this section, we present the research problems that will be specifically addressed in subsequent chapters, by introducing the related background and our proposed solutions.

1.3.1 Variable Selection in Point Process Modelling

Intensity estimation for point process models based on covariates has been extensively studied (e.g., Schoenberg, 2005; Waagepetersen, 2008; Guan & Shen, 2010; Guan, Jalilian & Waagepetersen, 2015). The classic approach is to consider a parametric intensity function that models the log-intensity as a linear combination of available covariates. Although this parametric model is easy to apply, existing studies worked on point pattern data with only a few covariates. When a larger number of environmental variables is available, directly incorporating them into the log-linear intensity function becomes problematic. To select the important variables, some studies relied on manual selection or simple correlation criteria. Recent research has explored regularized estimation techniques (Thurman & Zhu, 2014; Yue & Loh, 2015; Choiruddin, Coeurjolly &

Letué, 2018), however, they typically decided the model structure before fitting covariates, which may limit the use of covariate information and thus often requires additional modelling of residuals, i.e. with random effects. Moreover, due to potentially strong correlations among environmental variables, these variable selection methods can also perform unstably when handling a large number of putative covariates.

In contrast, machine learning approaches can detect the relationship between numerous variables and the response variable without prior information on model structure. Given specialized tools, they can mitigate the impact of correlations among candidate variables. A well-known technique in this domain is random forest, which relies on an ensemble of decision trees (Breiman, 2001). Each tree is trained on a subset of the data and the output is obtained by averaging the predictions across all trees. To assess the importance of a variable, one can permute its values over data observations and measure the mean increase of prediction error over the trees computed on the permuted data versus the original data. While machine learning methods can capture the complex covariate-response relationships, they are often difficult to interpret and are typically applicable for discrete data. Aggregating point pattern data into discrete areal unit data, however, can lead to information loss.

To tackle the variable selection challenge when fitting point pattern data with a large number of putative covariates, we propose a modelling procedure that combines the strengths of random forests with the rigorous framework of standard point process modelling for intensity estimation. We first employ random forests to identify the most important variables using aggregated areal unit data and then integrate the selected variables into a point process model to estimate the intensity based on original point pattern data. This approach is applied to the risk prediction of chimney fires, which exhibits obvious covariated-based phenomenon as introduced in Section 1.2.1. Beyond variable selection, the proposed method retains statistical interpretability, provides theoretical confidence intervals for intensity estimates, and enables second-order analysis for model validation.

1.3.2 Uncertainty Quantification under the Infill Regime

Uncertainty quantification plays a significant role in establishing comprehensive statistical inference for point process modelling. In the context of the log-linear parametric model for intensity estimation, the consistency and asymptotic normality have been studied for various estimators, including Poisson likelihood estimators (Schoenberg, 2005; Waagepetersen, 2008), logistic regression estimators (Baddeley et al., 2014; Choiruddin, Coeurjolly & Letué, 2018), and the quasi-likelihood approach (Guan, Jalilian & Waagepetersen, 2015; Xu, Waagepetersen & Guan, 2019). However, it is worth noting that most of the asymptotic results were derived under the increasing-domain regime where the observation window expands. In contrast, the literature focusing on the infill regime, where more points are observed in a fixed window, is relatively limited.

In chimney fire risk prediction, we opted for logistic regression estimators to

estimate the intensities of fire patterns. This choice is motivated by its straightforward implementation using standard software for generalized linear models and its effective use of a so-called dummy point process which allows for tuning so as to focus on important regions and times during model fitting. While the increasing-domain asymptotics can theoretically yield confidence intervals for fire predictions, the regime is unrealistic in practice. Firefighters point out that fire patterns often vary considerably from region to region. For example, chimney fires occurring in Twente, an eastern region of the Netherlands, exhibit clear, population-based spatial heterogeneity. In contrast, Amsterdam, the capital, shows a more stationary phenomenon, characterized by a lower risk due to the prevalence of newly constructed buildings of flats, hotels and stores where the use of chimneys is rare. In such scenarios, we obtain independent and identically distributed point patterns within fixed spatial domains across different time intervals, which was referred to as an infill asymptotic regime by Ripley (2012). Investigating the asymptotic properties of the intensity estimators under this regime is significant to construct reliable confidence intervals particularly for these cases.

To fill in the theoretical gap of infill asymptotics for intensity estimation in point process modelling, we derive the results of strong consistency and asymptotic normality for logistic regression estimators for the log-linear intensity function under the infill regime. Compared to related studies, we introduce easy-to-verify conditions and provide explicit, self-contained proofs. We also propose consistent estimators for the asymptotic covariance matrix. Furthermore, we extend the results to other estimators obtained from general unbiased estimating equations.

1.3.3 Flexible Nonparametric Intensity Estimation

The modelling procedure mentioned in Section 1.3.1 for estimating the intensity function of a point process over a large set of covariates has limitations, as variable selection and model fitting are performed at different levels of data aggregation. This separation may lead to inconsistencies because the important variables identified using areal unit data can differ from those identified based on point pattern data. In addition, the assumption of the log-linear parametric model is not always suitable in practice. In a follow-up study to predict the risk of kitchen fires, which also exhibits covariate-based phenomenon, Van Leeuwen (2022) applied the same variable selection method as for chimney fires, using random forests and conditional permutation importance techniques (cf., Section 2.7.2). Although key variables have been identified, it is challenging to formulate a similar, interpretable model structure, as many covariates were deemed significant. Therefore, it would be valuable to develop a fully integrated data-driven approach that combines variable selection and intensity estimation within a unified framework for point process modelling.

To this end, we explore nonparametric intensity estimation for point processes using covariates. Existing approaches include kernel intensity estimators

(e.g., Guan, 2008; Baddeley et al., 2012) and Gaussian Cox process models (e.g., Cunningham, Shenoy & Sahani, 2008; Kim, Asami & Toda, 2022). Kernel-based approaches apply standard kernel smoothing techniques to the covariate space, achieving computational efficiency but suffering from high local variability. Moreover, they are sensitive to bandwidth selection and are limited to data with no more than two covariates. Gaussian Cox process approaches model the latent intensity as a function of a Gaussian process and maximize the posterior probability function for model estimation. While performing better than kernel intensity estimators, they often require significant computational resources due to the complexity of approximating the intractable integration of the Gaussian process and their performances show a significant deterioration as the dimension of covariates increases. More recently, neural network-based point process models have been proposed by the machine learning community to capture the complex relationships between covariates and the intensity function in point processes (e.g., Okawa et al., 2019; Zhang, Kong & Zhou, 2023). Although powerful, these models typically require large data sets with millions of events for predictive performance. In classic statistics, however, point pattern data are often medium-sized, comprising hundreds or thousands of events, which emphasizes the need for an approach adept at handling a higher dimension of covariates when smaller data sets are available.

To address the challenge of flexibly estimating intensities for medium-sized point patterns in the presence of a large number of covariates, we build upon the concept introduced in Section 1.3.1. We propose a tree-based model to estimate the intensity function of a point process nonparametrically and design the associated learning algorithm and validation procedure. We also derive a dynamic weighted loss function, leveraging the boosting mechanism of XGBoost (Chen & Guestrin, 2016) over random forests, to improve the estimation efficiency for clustered point pattern data. We validate the superiority of our proposed method over existing approaches by comparing corresponding performance metrics in both numerical and real-world data experiments.

1.3.4 Bayesian Inference for Hierarchical Models

The hierarchical clustering model discussed in Section 1.2.3 has been well-studied in the literature (e.g., Van Lieshout, 1994; Møller & Waagepetersen, 2004). Among various structures, the most widely used one is the Neyman-Scott point process (Neyman & Scott, 1958), where both layers modelling the latent cluster centres and the actually observed events are assumed to be Poisson point processes. This model structure is quite popular due to the straightforward probability density of Poisson processes. For inference, frequentist methods (e.g., Waagepetersen, 2007; Mrkvička, Muška & Kubečka, 2012; Prokešová & Vedel Jensen, 2013) typically employ moment-based estimation, while Bayesian approaches (e.g., Van Lieshout & Baddeley, 2002; Møller & Waagepetersen, 2004) develop reversible jump Markov chain Monte Carlo algorithms to sample both state information of cluster centres and model parameters. Although

the Neyman-Scott process model is easy to build, it is not always suitable in real-world applications. First, the Poisson assumption for latent cluster centres may cause an overestimation of clusters when fitted to real data. Second, the Poisson assumption on the structure of a cluster is also restricted and may not adequately capture data complexity. Third, when the clustering evidence in data observations is weak, classic model fitting procedures can produce unstable and unreliable estimation results (Baddeley et al., 2022).

In arson fire risk prediction, a repulsive Strauss process (Strauss, 1975) is used to model cluster centres and a shifted Poisson distribution is assumed to govern the number of fire incidents in each cluster. These non-Poisson structures lead to an independent cluster point process, which shares the same framework of the Neyman-Scott process but allows both layers to be modelled by geneal point processes. This flexibility enables the independent cluster process model to introduce various interactions between cluster centres and accommodates a wide range of cluster structures, however, increases the difficulty of model inference even when only small deviations from the Poisson structure are added. For example, Møller & Torrisi (2005) derived explicit expressions of the second-order features of a general independent cluster point process. While the achieved formulas are theoretically insightful, they involve intractable high-dimensional integrals, making them challenging to solve and impractical for parameter estimation.

To solve the inference problem for general independent cluster point process models, we develop a two-step Bayesian approach. In the first step, we propose a Markov chain Monte Carlo method to sample the state information of cluster centres while holding the model parameters fixed. In the second step, we implement a Monte Carlo expectation-maximization algorithm to iteratively optimize model parameters, using the sampled states from the first step. Our inference approach allows for general point processes for both layers within the hierarchical framework, provided that their probability density functions have an analytical form with respect to the probability measure of a unit-rate Poisson process on the same observation window. Notably, compared to frequentist methods, Bayesian approaches offer the additional advantage of providing the density surface when estimating the locations of latent cluster centres.

1.4 Outline of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 presents the combined modelling procedure of machine learning and point process techniques to estimate intensities of point pattern data, storylined by the case study on chimney fire risk prediction. This chapter is based on the following two papers:

> Lu, C., Lieshout, M. N. M. van, Graaf, M. de & Visscher, P. (2021). Chimney fire prediction based on environmental variables. In *Proceedings of the 63rd ISI World Statis-*

tics Congress, 288–291. https://www.isi-web.org/publication/chimney-fire-prediction-based-environmental-variables.

Lu, C., Lieshout, M. N. M. van, Graaf, M. de & Visscher, P. (2023). Data-driven chimney fire risk prediction using machine learning and point process tools. *Annals of Applied Statistics*, 17(4), 3088–3111. https://doi.org/10.1214/23-AOAS1752.

Chapter 3 derives the theoretical asymptotic properties of logistic regression and other estimators obtained from general unbiased estimating equations for intensity estimation for spatio-temporal point processes under the infill regime. This chapter is based on the preprint:

Lieshout, M. N. M. van & Lu, C. (2022). Infill asymptotics for logistic regression estimators for spatio-temporal point processes. *arXiv:2208.12080*. https://doi.org/10.48550/arXiv.2208.12080.

Chapter 4 introduces the model, the learning algorithm and the applications of the fully data-driven, nonparametric tree-based estimation approach for point process intensity functions, based on the well-known gradient boosting mechanism, XGBoost. This chapter is based on the following two papers:

Lu, C., Guan, Y., Lieshout M. N. M. van & Xu, G. (2024). XGBoostPP: Tree-based estimation of point process intensity functions. *Journal of Computational and Graphical Statistics*, tentatively accepted. https://doi.org/10.48550/arXiv.2401.17966.

Lieshout M. N. M. van & Lu, C. (2024). M.N.M. van Lieshout and C. Lu's contribution to the discussion of 'the Discussion Meeting on probabilistic and statistical aspects of machine learning'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 86(2), 306–307. https://doi.org/10.1093/jrsssb/qkad150.

Chapter 5 proposes the two-step Bayesian inference framework to estimate general independent cluster point processes and validates the approach on a specific model structure designed for predicting car arson fire risk. It will also be submitted as a preprint.

Chapter 6 concludes the thesis and gives directions for future work.

CHAPTER 2

Data-driven Chimney Fire Risk Prediction Using Machine Learning and Point Process Tools

2.1 Introduction

During the last decade, the Dutch fire and rescue services have been developing an interest in applying business intelligence to improve their strategy of fire prediction and prevention (NVBR, 2010). To prepare for risk reducing measures, such as essential public awareness campaigns and proper fire staffing and equipment arrangements, accurate predictions are required. In this chapter, we focus on chimney fires, as they occur frequently, rely heavily on environmental factors and impact people's daily life. Collaborating with the Twente Fire Brigade, we conduct a complete risk prediction study for chimney fires. We define fire risk prediction as an occurrence modelling problem, analyze underlying patterns and design appropriate prediction models. Our approach for chimney fire prediction is general and can be transferred to similar fire patterns.

The literature for fire risk prediction is mostly concerned with wildfires. Overall, the prediction approaches can be divided into two categories: machine learning based approaches (e.g., Rodrigues & De la Riva, 2014; Jain et al., 2020; Malik et al., 2021) and statistical approaches (e.g., Turner, 2009; Møller & Díaz-Avalos, 2010; Xu & Schoenberg, 2011). Usually, machine learning based approaches do not require prior knowledge but can detect the dependence between fire risk and a large number of environmental variables automatically using specialized learning algorithms, such as logistic regression (Preisler et al., 2004), support vector machine (Sakr et al., 2010), decision tree (Stojanova et al., 2012), random forest (Rodrigues & De la Riva, 2014) and neural network (Satir, Berberoglu & Donmez, 2016). Most machine learning algorithms are applicable for discrete data, whereas hazard maps of fire risk are continuous. Moreover, fire occurrences are usually recorded as spatio-temporal point patterns. To resolve such discrepancies, certain studies (e.g., Sakr et al., 2010; Stojanova et al., 2012; Satir, Berberoglu & Donmez, 2016) discretized fire incidents into areal unit data and transformed risk prediction from an occurrence modelling problem to a risk scale classification problem, i.e. labelling different amounts of fires to corresponding

risk scales. However, doing so may result in a loss of information. In addition, machine learning approaches often require relatively large amounts of data to obtain a satisfactory performance and, due to their 'black box' behaviour, it is difficult to interpret the influence of an explanatory variable on fire risk in terms that are meaningful for practical implementation for fire services.

In contrast, statistical approaches enable to learn spatio-temporal point patterns directly. Moreover, they are interpretable based on specific, parametric, mathematical models and allow for theoretical confidence intervals. For instance, Hering, Bell and Genton (2009) and Costafreda-Aumedes, Comas & Vega-Garcia (2016) used the K-function to analyze clustering patterns of wildfire data. Mandallaz & Ye (2011) and Boubeta et al. (2015) employed Poisson structures to generate hazard maps of fire occurrences. Møller & Díaz-Avalos (2010) considered spatial and temporal explanatory variables in a shot-noise Cox process and fitted it with minimum contrast techniques. Pereira et al. (2013) and Serra et al. (2014) modelled fire occurrences using log-Gaussian Cox processes, which were particularly designed to simulate latent phenomena. Recently, a Bayesian framework was developed and suggested to improve fire prediction (Verdoy, 2019). Pimont et al. (2020) also employed a log-Gaussian Cox process model to predict wildfires but established Bayesian inference for model components using integrated nested Laplace approximation (Rue, Martino & Chopin, 2009). Koh et al. (2023) developed a joint hierarchical model framework by combining extreme-value theory and point processes and studied summer wildfire data for the French Mediterranean basin. For chimney fires, School (2018) proposed a log-Gaussian Cox process model to predict the risk based on explanatory variables selected by Pearson correlation coefficients and random effects simulated by Gaussian random fields. Most statistical studies (e.g., Møller & Díaz-Avalos, 2010; Pereira et al., 2013) first determined their model structures and selected and fitted explanatory variables afterwards. This procedure does not fully exploit information contained in variables and sometimes requires complex modelling of residuals. In addition, the selection of explanatory variables is conducted either manually (Turner, 2009) or using basic statistical methods (Yang et al., 2015; School, 2018). Some studies (Thurman & Zhu, 2014; Yue & Loh, 2015; Choiruddin, Coeurjolly & Letué, 2018) applied regularized penalty functions in the estimation of point processes for variable selection. However, a parametric form of the intensity function including all variables needs to be specified in advance.

To conduct a data-driven fire risk prediction study, we combine machine learning and point process tools in our modelling procedure to leverage the advantages of both types of approaches. Specifically, we use the permutation importance techniques of random forests to select important variables, as they are recommended as relatively accurate machine learning methods for fire prediction (Rodrigues & De la Riva, 2014). In a pilot study, cf. Section 2.4.6, we designed a generalized linear Poisson model based on the selected variables and fitted it on areal unit data to predict the chimney fire risk in Twente. In this chapter, we complete the theoretical background and implementation details for the variable

selection part and refine the areal unit model to a spatio-temporal Poisson point process model. Since point patterns (cf., Figure 2.1 (right)) are targeted directly, the new model does not depend on the scale of the areal units, leads to a continuous hazard map for fire prediction and enables more detailed data analyses, such as point interaction tests.

Specifically, our contributions are as discussed below. First, we propose a modelling procedure for fire prediction, which is fully data-driven and, meanwhile, allows for model interpretability: (i) we use machine learning algorithms in a supporting role to select explanatory variables for fire risk nonparametrically, so that variable information can be exploited; (ii) we adhere to a statistical Poisson point process model for fire occurrences with parametric structures designed on the selected explanatory variables, so that complete model inference can be conducted. Second, in the fitting of the point process model, we efficiently estimate the parameters using logistic regression estimation (Baddeley et al., 2014) by tuning it to focus on important parts in space and time and provide theoretical confidence intervals.

The remainder of this chapter is organized as follows. Section 2.2 introduces the chimney fire data and some data pre-processing steps. Section 2.3 elaborates on the selection of important variables. In Section 2.4, we motivate and fit a Poisson point process model and compare it to the areal unit model proposed in the pilot study. Section 2.5 tests for point interactions in the fire data and validates the Poisson model assumption of absence of such interactions. Section 2.6 reviews our modelling procedure and explores the role of the dummy intensity in logistic regression estimation. In Section 2.7, we present extended applications of our approach to other fire prediction case studies. Section 2.8 provides a summary and some ideas on future work.

2.2 Data

In this section, we introduce the chimney fire data, the relevant environmental variables and the data pre-processing operations.

2.2.1 Chimney Fires and Environmental Variables

Collaborating with the Twente Fire Brigade, we collected the data of all reported chimney fire incidents occurring between January 1, 2004 and December 31, 2020, in the Twente region, in the eastern part of the Netherlands (map shown in Figure 2.1 (left)). After a manual check to remove obvious mistakes, the data set consists of 1759 incidents. Each incident is reported individually with its ID number, location, time and a brief description of the circumstances of the fire and rescue processes. Locations are documented using Dutch RD coordinates in the unit of metre, while time is recorded in the unit of day to match the weather data. The spatial and temporal projections are plotted, respectively, in Figure 2.1 (right) and Figure 2.2. It is clearly visible in Figure 2.1(right) that

the spatial distribution of chimney fires is heterogeneous in the sense that most incidents occur in urban areas, especially in cities with a higher population such as Almelo, Hengelo and Enschede (cf., Figure 2.1(left)). Apparent clustering may also arise, because neighbouring buildings tend to have identical chimney types. Moreover, locations neighbouring in space and time share similar weather conditions. In Figure 2.2, the temporal distribution of chimney fires is periodic: chimney fires occur more frequently in winter than in summer.

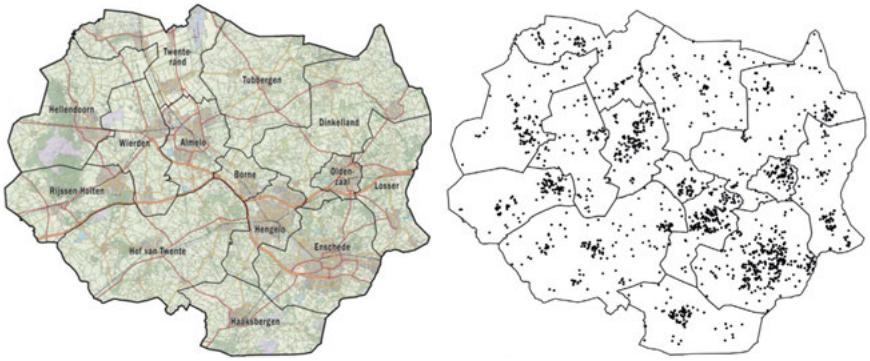


Figure 2.1 Map of Twente municipalities (left) and spatial projection of the chimney fire incidents during 2004–2020 (right).

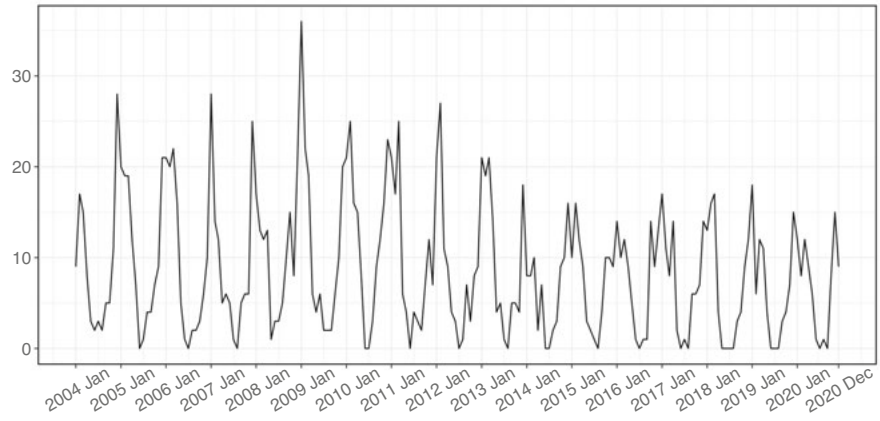


Figure 2.2 Temporal projection (monthly counts) of the chimney fire incidents during 2004–2020.

According to the experts from the Twente Fire Brigade, various environmental factors may influence the risk of chimney fires. Spatially, building types, population density and compositions and urbanity degrees in an area may determine the baseline fire risk. Temporally, season and weather conditions influence the risk as well. Other factors, such as fueling sources, chimney types and dominant

human causes, may also affect it, however, their information is not accessible. Table 2.1 lists 27 putative explanatory variables with specific abbreviations, descriptions and sources. Among them, population and urbanity information are recorded over 6291 pre-defined $500m \times 500m$ area boxes in Twente, whereas building information contains the precise locations, ages and functions of houses. Temporal data comes in the form of daily weather observations from the weather station (Twenthe) at Twente airport. To assess the influence of small variations in weather among different parts of the Twente region, we also collected the data from two neighbouring weather stations (Heino and Hupsel) outside Twente for analysis. Their locations can be found on `https://www.knmi.nl`.

2.2.2 Data Pre-processing

To unify the data to be used in our modelling procedure, we perform several data pre-processing operations on environmental variables.

First, we further extract building information at the level of living units, as it may reflect the geographical information of chimneys in a more representative way. For instance, two families living in one semi-detached house actually use their individual chimneys. Second, some spatial variables (i.e., population and urbanity information) consist of a list of historical data, whereas others (i.e., building information) are only accessed in a single actual value. To enable similar treatment of the data, we use averaged values over the time period of interest for those variables consisting of a list of historical data, so that for all spatial variables, a single value is accessible. Moreover, we only consider the buildings that are currently in use and assume that buildings keep their functions and types during the time period of interest. Observing that the age and type information of some buildings are missing, we assign them the label 'extra' during the counting process. Finally, we exclude the fire incidents and weather data on the leap days in the leap years, while in future fire risk prediction, we will use the prediction of February 28 to compensate for the missing prediction of the leap day, February 29.

In addition, particularly for the point process modelling phase in Section 2.4, in order to obtain the data that records spatial variables for every location in Twente (i.e., density maps) from either areal unit data (i.e., population and urbanity information) or data with precise coordinates (i.e., building information), we employ kernel smoothing. For temporal variables, we assume that their values at different times of a day remain invariant and equal to daily mean values. Recalling Table 2.1, we then introduce the following notation for environmental variables:

- $V_{\sigma,i}(\boldsymbol{w})$: the smoothed value of the i-th spatial variable at location $\boldsymbol{w}$,
- $V_{\tau,i}(t)$: the assigned value of the i-th temporal variable at time t,

where $(\boldsymbol{w}, t)$ denotes any location and time combination in the spatio-temporal domain.

It is worth noting that 'Urbanity' is a categorical variable, where the urbanity of a neighbourhood is evaluated into levels varying from 1 to 5 (urban–not urban), thus will be treated as numerical in random forests. 'Visibility' is also

Table 2.1 Putative explanatory variables, with their abbreviations, descriptions and sources. σ: spatial variables, τ: temporal variables. $V_{\sigma,1}$–$V_{\sigma,11}$ are collected from NIPV: *Nederlands Instituut Publieke Veiligheid*; $V_{\sigma,12}$–$V_{\sigma,22}$ are collected from CBS: *Centraal Bureau voor de Statistiek*; $V_{\tau,1}$–$V_{\tau,5}$ are collected from KNMI: *Koninklijk Nederlands Meteorologisch Instituut*.

Variable	Abbrev	Description
$V_{\sigma,1}$	House	The total number of houses
$V_{\sigma,2}$	House_indu	The number of houses with an industrial function
$V_{\sigma,3}$	House_hotl	The number of houses with a hotel function
$V_{\sigma,4}$	House_resi	The number of houses with a residential function
$V_{\sigma,5}$	House_20	The number of houses constructed before 1920
$V_{\sigma,6}$	House_2045	The number of houses constructed between 1920 and 1945
$V_{\sigma,7}$	House_4570	The number of houses constructed between 1945 and 1970
$V_{\sigma,8}$	House_7080	The number of houses constructed between 1970 and 1980
$V_{\sigma,9}$	House_8090	The number of houses constructed between 1980 and 1990
$V_{\sigma,10}$	House_90	The number of houses constructed after 1990
$V_{\sigma,11}$	House_frsd	The number of free standing (detached or semi-detached) houses
$V_{\sigma,12}$	Resid	The total number of residents
$V_{\sigma,13}$	Resid_14	The number of residents with an age in the range of 0 till 14
$V_{\sigma,14}$	Resid_1524	The number of residents with an age in the range of 15 till 24
$V_{\sigma,15}$	Resid_2544	The number of residents with an age in the range of 25 till 44
$V_{\sigma,16}$	Resid_4564	The number of residents with an age in the range of 45 till 64
$V_{\sigma,17}$	Resid_65	The number of residents with an age of 65 or higher
$V_{\sigma,18}$	Man	The number of male residents
$V_{\sigma,19}$	Woman	The number of female residents
$V_{\sigma,20}$	Address	The number of addresses in the neighbourhood
$V_{\sigma,21}$	Urbanity	The urbanity of the neighbourhood
$V_{\sigma,22}$	Town	Boolean variable indicating the presence of a town
$V_{\tau,1}$	WindSpeed	Daily mean wind speed (km/h)
$V_{\tau,2}$	Temperature	Daily mean temperature (°C)
$V_{\tau,3}$	WindChill	Daily mean wind chill (°C) (calculated from $V_{\tau,1}, V_{\tau,2}$)
$V_{\tau,4}$	Sunshine	Daily sunshine duration (h)
$V_{\tau,5}$	Visibility	Daily minimum visibility

a categorical variable, where the minimum visibility distances (0–∞ km) are defined to levels varying from 1 to 80, thus will be treated as numerical in random forests as well.

2.3 Selection of Explanatory Variables

Obviously, not all environmental variables listed in Table 2.1 are required to model the risk of chimney fires and some of them are mutually dependent. To select the most informative ones, we perform a nonparametric variable importance analysis using the permutation importance techniques of random forests (Breiman, 2001). For completeness, we elaborate here on both the theoretical background and implementation details.

2.3.1 Random Forests and Permutation Importance

Random forests (Breiman, 2001) are widely used as robust classification and regression methods in many applications, such as risk prediction (Wongvibulsin, Wu & Zeger, 2019) and data mining (Schonlau & Zou, 2020). A random forest is usually composed of hundreds or thousands of decision trees, where each tree is generated on a sampled subset of the data by repeated bagging (i.e., bootstrap sampling with replacement) and trained to fit explanatory variables to the response variable. Afterwards, a combined result over all trees will be reported as the final output. In addition, random forests can also be used to assess the importance of a variable by means of permutation importance techniques (Breiman, 2001; Strobl & Zeileis, 2008; Altmann et al., 2010). Through randomly permuting the values of a variable over the observations, the importance of a variable is defined as the mean increase of prediction error over all trees computed on the permuted data compared to that computed on the original data.

More formally, consider a regression problem on a dataset $\mathcal{D}$ with n observations. Suppose that, for each observation in $\mathcal{D}$, there are m explanatory variables. Let x_i^j and y_i, with $i = 1, ..., n$ and $j = 1, ..., m$, denote the j-th explanatory variable and the response variable, respectively, for the i-th observation; $\boldsymbol{x}_i$ is the vector that collects all x_i^j. The construction of a random forest consists of the generation of E decision trees. To generate a tree e, a subset of $\mathcal{D}$, denoted as $\mathcal{B}(e)$, is sampled by bagging. At each node of the tree, a number of explanatory variables are selected randomly from all variables as the candidates. Then, one of the candidates, say x^j, is used to split the node into two subsets, $\mathcal{B}_L(e)$ and $\mathcal{B}_R(e)$ (e.g., for a numerical x^j, $\mathcal{B}_L(e) = \{(\boldsymbol{x}_i, y_i) : x_i^j \leq c; (\boldsymbol{x}_i, y_i) \in \mathcal{B}(e)\}$ and $\mathcal{B}_R(e) = \{(\boldsymbol{x}_i, y_i) : x_i^j > c; (\boldsymbol{x}_i, y_i) \in \mathcal{B}(e)\}$), in such a way that the residual sum of squares

$$\sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{B}_L(e)} \left[y_i - \bar{y}_{\mathcal{B}_L(e)}\right]^2 + \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{B}_R(e)} \left[y_i - \bar{y}_{\mathcal{B}_R(e)}\right]^2 \tag{2.1}$$

is minimized. Here, $\bar{y}_{\mathcal{B}_L(e)}$ and $\bar{y}_{\mathcal{B}_R(e)}$ denote the mean of the response variables in corresponding subsets. Such a node splitting procedure is continued until tree e satisfies certain preconditions, e.g. reaching the maximum number of levels of a tree. Similar tree generation processes are employed to construct all other trees in the forest. To measure the importance of an explanatory variable, again say x^j, in tree e, we randomly permute its values over the out-of-bag observations of the tree, denoted as $\mathcal{OB}(e)$ (i.e., $\mathcal{OB}(e) = \mathcal{D} \setminus \mathcal{B}(e)$). Write π_j for the permutation. The value of the j-th explanatory variable for the response y_i for an observation $(\boldsymbol{x}_i, y_i)$ then changes from x_i^j to $x_{\pi_j(i)}^j$ with $(\boldsymbol{x}_i, y_i) \in \mathcal{OB}(e)$, while the values of other explanatory variables are left unchanged. Thus, the increase of prediction error is

$$I\left(x^j; e\right) = \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{OB}(e)} \frac{\left(y_i - \widehat{y}_{i,\pi_j}\right)^2}{|\mathcal{OB}(e)|} - \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{OB}(e)} \frac{\left(y_i - \widehat{y}_i\right)^2}{|\mathcal{OB}(e)|}, \tag{2.2}$$

where $\widehat{y}_i$ denotes the prediction in tree e for observation $(\boldsymbol{x}_i, y_i)$ using original explanatory variables, $\widehat{y}_{i,\pi_j}$ denotes the corresponding prediction but using explanatory variables with the j-th variable permuted and $|\mathcal{OB}(e)|$ denotes the number of out-of-bag observations for the tree. Finally, the mean increase of the prediction error over all trees, $\sum_e I(x^j; e)/E$, is used to illustrate the importance of the j-th variable, x^j, in the forest.

In practice, the original construction algorithms of random forests tend to bias the variable selection at tree nodes to factorial variables with many categories or continuous variables with many cut points. Moreover, traditional permutation importance techniques can sometimes be misled by correlated explanatory variables. To address these problems, unbiased random forests (Hothorn, Hornik & Zeileis, 2006; Strobl et al., 2007) and conditional permutation importance techniques (Strobl et al., 2008) were proposed based on a conditional inference framework of recursive partitioning. Compared to earlier methods, the idea is to partition the variable space in order to obtain groups of observations with similar association patterns instead of groups of observations with merely similar values of the response variable. For instance, to measure the importance of variable x^j under conditional permutation in tree e, the set of out-of-bag observations, $\mathcal{OB}(e)$, is first partitioned into a grid where each block shares the same information on the remaining variables, $\boldsymbol{x} \setminus x^j$. Then, the same permutation operation as introduced above is applied to each block and the increase of the prediction error is computed and summed up over all blocks to illustrate the importance of x^j. With significant increments and optimizations, the random forest approach has been very useful to measure variable importance nonparametrically considering variable dependence (Strobl, Hothorn & Zeileis, 2009).

Note that the notation used in this section is only valid here to explain random forests and permutation importance techniques statistically.

2.3.2 Variable Importance Analysis

To measure the importance of spatial and temporal variables, we implement separated experiments in space and time on areal unit data (cf., Section 2.4.6). The reason for this is that, in the original data, spatial variables are available as a single actual value for every area box while temporal variables are available as historical time series.

Specifically, for spatial variables, we group the incidents by the 6291 predefined $500m \times 500m$ area boxes and merge them into a large data frame consisting of 6291 rows and 23 columns. Each row indicates an area box and the 23 columns refer to the number of incidents occurring in that box (response variable) as well as the values of 22 spatial environmental variables. We find that some boundary area boxes lie only partially in Twente. However, the population statistics for such boxes do not distinguish between Twente and the neighbouring regions. To avoid bias, we filter out the data of these boundary boxes in our variable importance analysis. For temporal variables, we group the incidents in 6205 days and merge them into a data frame consisting of 6205 rows and 6

columns. Each row indicates a specific day and the 6 columns refer to the number of incidents on that day (response variable) as well as the values of 5 temporal environmental variables.

In the analysis, we construct two unbiased random forests of 2000 trees to fit spatial and temporal variables to the number of incidents in corresponding data tables respectively and set the proportion of the number of input variables that are randomly sampled as the candidates at a tree node to approximately 1/3 (i.e., spatial: 7 candidates, temporal: 2 candidates). Considering high correlations between certain variables (e.g., population and urbanity information), we use conditional permutation importance techniques (Strobl et al., 2008) instead of the traditional ones (Breiman, 2001) to suppress the importance scores biasing towards correlated variables. Our implementation of variable importance analysis is based on the R-package *party* (Hothorn et al., 2006; Strobl et al., 2007). We use the method proposed in Debeer & Strobl (2020) to compute the conditional permutation importance, which provides a faster computation and shows more stable results than the original implementation (Strobl et al., 2008).

The variable importance results under conditional permutation for spatial and temporal variables are plotted in Figure 2.3 and Figure 2.4, respectively. Note that the importance scores for spatial and temporal variables are not comparable, as the two random forests are fitted on spatial and temporal variables, separately. For comparison, we also plot the results under traditional permutation. In all plots, the y-axis refers to the increase of the squared prediction error when traditional or conditional permutation on a variable is applied. A large increase indicates that the variable is very important for correct predictions, while a decrease indicates that the variable has no influence on or even hampers the prediction. If we compare the results under traditional and conditional permutation, we see that the bias caused by correlated variables is suppressed well. For instance, the number of residents aged between 45 and 64 and the total number of residents have the largest importance scores under traditional permutation, with the number of free standing houses coming third. However, the latter has the highest variable importance score under conditional permutation. A possible explanation could be that free standing houses are mostly occupied by families whose adult members are aged between 45 and 64, which implies a strong positive correlation among all three categories. Moreover, other age groups such as the elderly or families with young children may be less inclined to use their chimneys. The importance scores of the two variables concerning residents decrease a lot under conditional permutation. Similar observations hold for temporal variables as well. Both wind chill and temperature obtain high importance scores under traditional permutation although they are correlated. Conditional permutation detects the underlying correlation and suppresses temperature, as wind chill is defined in terms of both temperature and wind speed information. However, since wind speed still obtains a relatively large importance score even under conditional permutation, we will further consider it in the modelling phase. The additional analysis to assess the influence of small weather variations among dif-

ferent parts in Twente indicates that these variations can be ignored (detailed results are provided in Appendix 2.9).

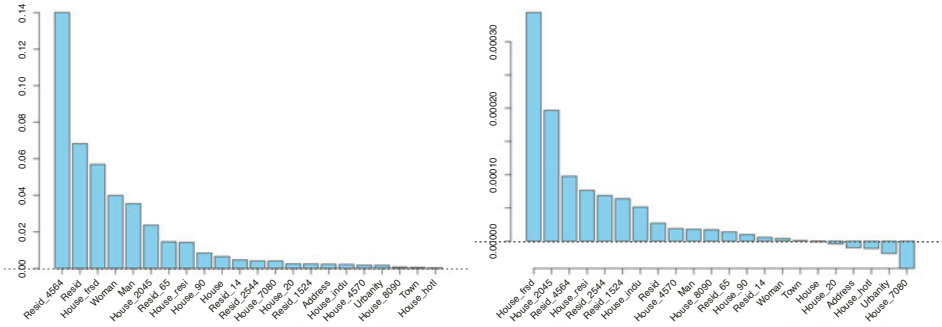


Figure 2.3 Importance obtained for spatial variables using traditional (left) and conditional (right) permutation techniques.

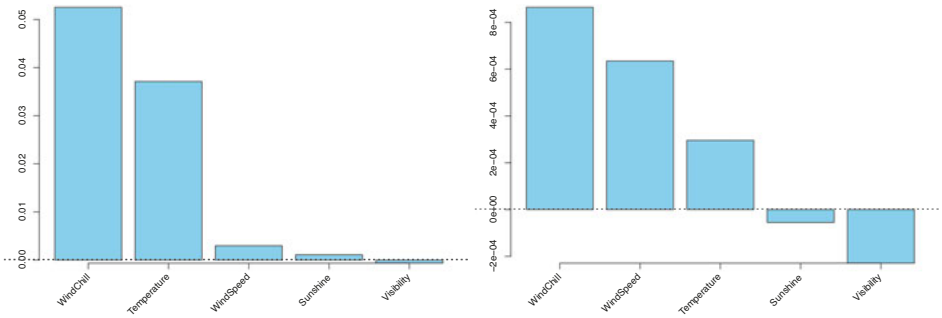


Figure 2.4 Importance obtained for temporal variables using traditional (left) and conditional (right) permutation techniques.

Overall, according to the variable importance under conditional permutation, the number of buildings constructed between 1920 and 1945 and the number of free standing houses are the most important spatial variables, and wind chill and wind speed are the most important temporal variables. The results could have the following explanations: (i) most free standing houses contain chimneys whereas other types do not; (ii) chimney pipes in old buildings tend to be made of brick rather than metal, which increases the risk to catch fires; (iii) strong wind can fuel a fire; (iv) wind chill reflects people feeling cold thus inducing them to use their fires and chimneys.

2.4 Poisson Point Process Model

In this section, we develop a spatio-temporal point process model for chimney fire prediction. First, we discuss the motivation for a nested model structure based on the selected variables from Section 2.3. Second, we define a Poisson point process model using this structure. Afterwards, we present model fitting and selection procedures and propose theoretical confidence intervals for both model parameters and predicted fire intensities. Finally, we compare the point process model to the areal unit model developed in our pilot study and illustrate the former's advantages.

2.4.1 Model Motivation

In preparation for an appropriate model structure for chimney fire prediction, we perform a preliminary investigation on the relations between the selected explanatory variables and chimney fire occurrences using areal unit data.

Recalling the selected variables in Section 2.3, we divide the houses in Twente into four house types depending on their ages (i.e., whether they have been constructed between 1920 and 1945 or not) and on whether they are free standing or not and plot the monthly intensities of chimney fires per house for different house types separately. Figure 2.5 shows that, generally, chimney fire occurrences in all house types are periodic, with incidents concentrated in the colder seasons. However, different house types run different risks of chimney fires: the intensities for old houses (i.e., House_2045) and free standing houses are higher than others. Moreover, note that the patterns are not perfectly periodic; the amplitudes of the peaks vary per year. To model such varying patterns, we need to take wind chill and wind speed into account.

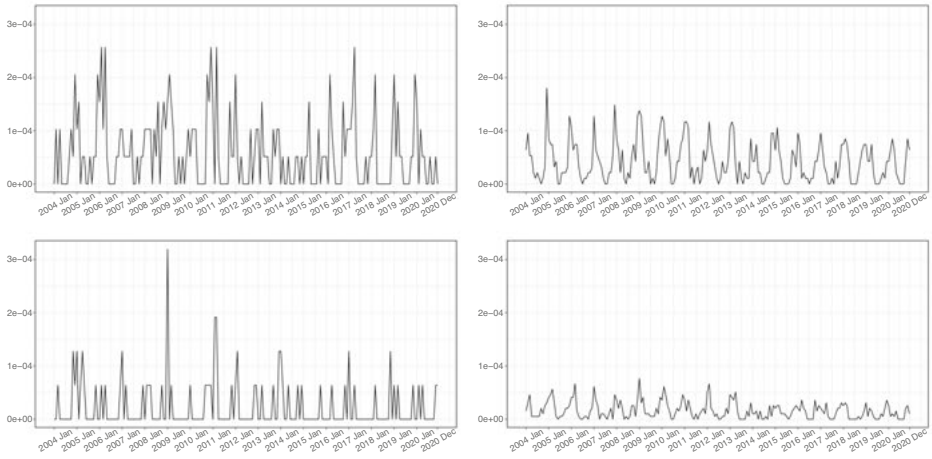


Figure 2.5 Monthly intensities of chimney fires per house for different house types during 2004–2020.

In addition, we plot the numerical relations between chimney fire occurrences and the number of houses of different types in Figure 2.6. To approximate the underlying trends in the scatter plots, we apply the locally estimated scatter plot smoothing method (Cleveland, Grosse & Shyu, 1992) and plot the estimated trends. It is interesting to find that, for each house type, chimney fire occurrences increase approximately linearly in the number of houses of that type. However, this observation only holds when the number of houses is not very large. If it is too large, the number of chimney fires tends to a saturated value.

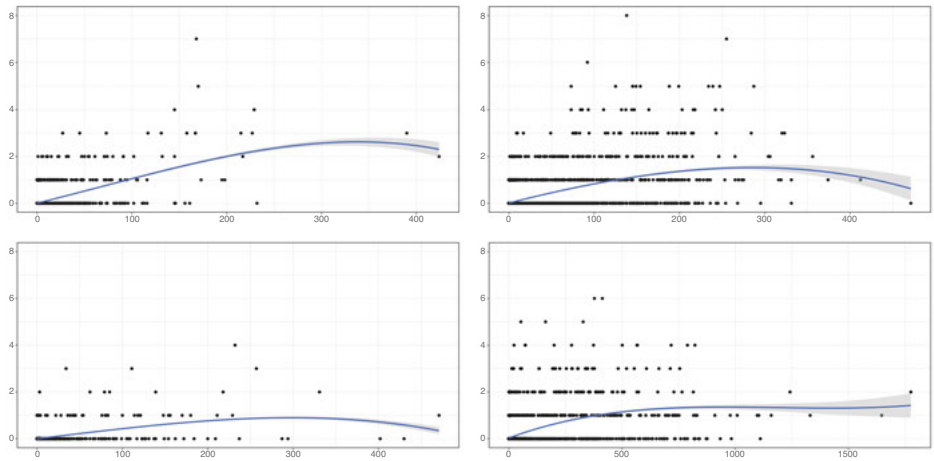


Figure 2.6 Numerical relations between chimney fire occurrences (y-axis) and the number of houses of different types (x-axis). The blue curves are the estimated trends using the locally estimated scatterplot smoothing method with parameter $span = 1$. The 95% envelopes are also included to show the confidence of such estimates.

To summarize, we draw the following conclusions: (i) chimneys in different house types catch fires at a type-dependent rate, and the rate is influenced by both seasonal information and temporal explanatory variables; (ii) chimney fire occurrences in the houses of a specific type are approximately proportional to the number of houses of that type.

2.4.2 Model Structure

Based on the motivation above, we design a nested Poisson model structure for chimney fire risk prediction. A model using this structure for fitting areal unit data will be introduced in Section 2.4.6. Here, we construct a point process model which considers the same structure but learns point patterns (cf., Figure 2.1(right)) directly. We assume that the explanatory variables for the areal unit model and the point process model are identical.

Suppose that each house catches chimney fires independently at a type-dependent rate that varies in time. Under this assumption, the fire predic-

tion model is a spatio-temporal Poisson point process defined on the Twente region and the period 2004–2020 with an intensity function of the form $\lambda(\boldsymbol{w}, t) = \sum_k \lambda_k(\boldsymbol{w}, t)$, where $\lambda(\boldsymbol{w}, t)$ indicates the overall fire risk at location $\boldsymbol{w}$ at time t and $\lambda_k(\boldsymbol{w}, t)$ indicates the risk at location $\boldsymbol{w}$ at time t for the houses of type k. Furthermore, we define $\lambda_k(\boldsymbol{w}, t) = h_k(\boldsymbol{w})\phi_k(t)$, where $h_k(\boldsymbol{w})$ indicates the density of house type k at location $\boldsymbol{w}$ and $\phi_k(t)$ indicates the risk for a house of that type at time t. Such a structure reflects the conclusions reached in Section 2.4.1, although it relies on the condition that the density of houses at a location is not too large (cf., Figure 2.6).

More formally, our nested Poisson point process model is defined by its intensity function

$$\Lambda(\boldsymbol{w}, t) = \sum_{k=1}^{4} \lambda_k(\boldsymbol{w}, t) = \sum_{k=1}^{4} h_k(\boldsymbol{w})\phi_k(t). \tag{2.3}$$

The density maps of houses of four types, $h_k(\boldsymbol{w})$, as shown in Figure 2.7, are derived from corresponding building information (i.e. $V_{\sigma,6}(\boldsymbol{u})$ and $V_{\sigma,11}(\boldsymbol{u})$ in Table 2.1), by smoothing using Gaussian kernels with a standard deviation of 1000 metres. This standard derivation is suggested by the experts from the Twente Fire Brigade, considering the distance between cities in Twente. Moreover, since both seasonal information and temporal variables, wind speed and wind chill (i.e. $V_{\tau,1}(t)$ and $V_{\tau,3}(t)$ in Table 2.1) are taken into account, we propose a temporal intensity function $\phi_k(t)$ for houses of type k of the form

$$\begin{aligned}\phi_k(t) = \exp\big[&\text{Harmonic}(t; o_{k,1}) + \text{Polynom}(V_{\tau,1}(t); o_{k,2}) + \\ &\text{Polynom}(V_{\tau,3}(t); o_{k,3}) + \text{Polynom}(V_{\tau,1}(t)V_{\tau,3}(t); o_{k,4})\big],\end{aligned} \tag{2.4}$$

where a harmonic function (cosine and sine functions) with order $o_{k,1}$ is employed to model seasonal variations and two polynomial functions with order $o_{k,2}$ and $o_{k,3}$ are used to model information of wind speed and wind chill. In addition, we employ another polynomial function of $V_{\tau,1}(t)V_{\tau,3}(t)$ with order $o_{k,4}$ to allow for interactions between wind speed and wind chill. The exponential function is applied to guarantee that the intensity function stays positive. For purposes of parametric representation, the temporal intensity function $\phi_k(t)$ can also be described as $\phi_k(t; \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ indicates the vector of coefficients in the harmonic and polynomial functions for house type k. Accordingly, $\lambda_k(\boldsymbol{w}, t) = \lambda_k(\boldsymbol{w}, t; \boldsymbol{\theta}_k)$.

2.4.3 Model Fitting

In principle, we could apply maximum likelihood estimation here to maximize the log-likelihood function of our Poisson point process model which reads

$$\sum_{k=1}^{4} \left[\sum_{\boldsymbol{x}_k} \log \lambda_k(\boldsymbol{x}_k; \boldsymbol{\theta}_k) - \int_{\text{Twente}} \int_{\text{Year 2004–2020}} \lambda_k(\boldsymbol{w}, t; \boldsymbol{\theta}_k) \mathrm{d}\boldsymbol{w}\mathrm{d}t \right], \tag{2.5}$$

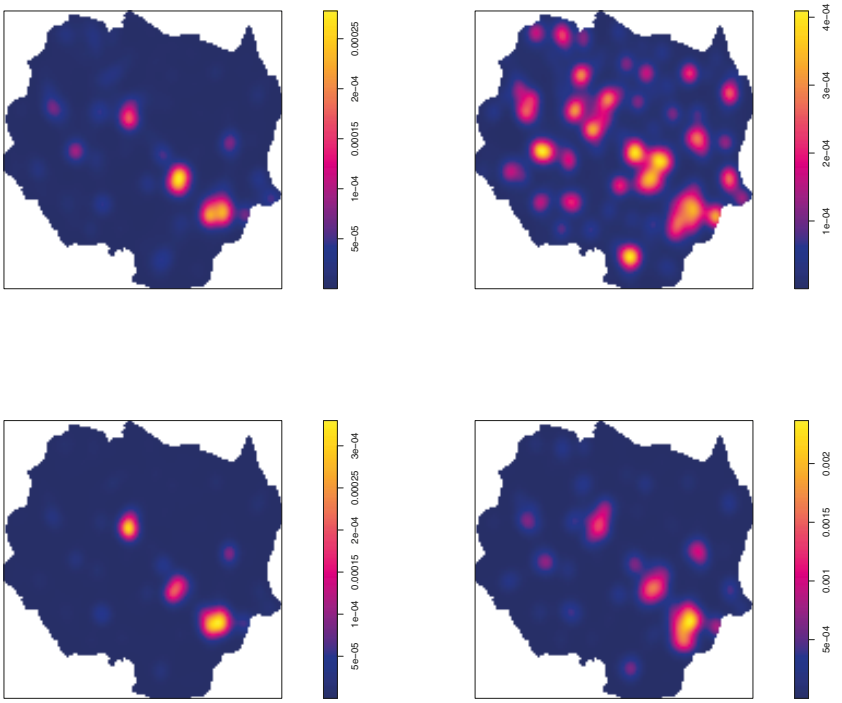


Figure 2.7 Density maps of houses of four types. The unit in all plots is metre^{-2}. Note that the intensity bars have different scales.

where $\boldsymbol{x}_k$ denotes combinations of location and time and runs through chimney fire incidents occurring in houses of type k. However, since we have a continuous space-time domain, the integral in the log-likelihood function must be approximated numerically, e.g. using quadrature points (Baddeley & Turner, 2000). In our case, the number of quadrature points would have to be rather large, thus, we prefer to use logistic regression estimation (Baddeley et al., 2014) to fit the Poisson point process model, where we can freely adjust the distribution of quadrature points.

The idea behind the logistic regression estimation approach is the well-known Campbell-Mecke theorem (see, e.g., Daley & Vere-Jones, 2009). Consider a point process X that is defined on a space-time domain $\mathcal{W} \times \mathcal{T} \subset \mathbb{R}^2 \times \mathbb{R}$ with intensity function $\lambda(\boldsymbol{w}, t)$. For any vector of real-valued non-negative or integrable

functions $\boldsymbol{f}(\boldsymbol{w}, t)$ defined on $\mathcal{W} \times \mathcal{T}$, the Campbell-Mecke theorem reads

$$\mathbb{E}\left[\sum_{\boldsymbol{x} \in X} \boldsymbol{f}(\boldsymbol{x})\right] = \int_{\mathcal{W}} \int_{\mathcal{T}} \boldsymbol{f}(\boldsymbol{w}, t) \lambda(\boldsymbol{w}, t) \mathrm{d}\boldsymbol{w} \mathrm{d}t, \tag{2.6}$$

where x runs through the points of X. Suppose that the intensity function λ is parametric and depends on a parameter vector $\boldsymbol{\theta}$, thus can be described as $\lambda(\boldsymbol{w}, t; \boldsymbol{\theta})$. One may estimate both sides of the Campbell–Mecke theorem and then equate them to obtain estimating equations for $\boldsymbol{\theta}$.

The logistic regression estimation approach is based on the components of the vector function

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{w}, t) &= \frac{\partial}{\partial \boldsymbol{\theta}} \log \left[\frac{\lambda(\boldsymbol{w}, t; \boldsymbol{\theta})}{\lambda(\boldsymbol{w}, t; \boldsymbol{\theta}) + \rho(\boldsymbol{w}, t)}\right] \\ &= \frac{\rho(\boldsymbol{w}, t)/\lambda(\boldsymbol{w}, t; \boldsymbol{\theta})}{\lambda(\boldsymbol{w}, t; \boldsymbol{\theta}) + \rho(\boldsymbol{w}, t)} \nabla \lambda(\boldsymbol{w}, t; \boldsymbol{\theta}), \end{aligned} \tag{2.7}$$

where $\nabla \lambda(\boldsymbol{w}, t; \boldsymbol{\theta})$ denotes the gradient of $\lambda(\boldsymbol{w}, t; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and $\rho(\boldsymbol{w}, t)$ is a positive-valued function defined on $\mathcal{W} \times \mathcal{T}$. Here, one needs to assume that $\lambda > 0$. To estimate the integral in (2.6) with $\boldsymbol{f}(\boldsymbol{w}, t)$ defined in (2.7), one can use a 'dummy' point process D on $\mathcal{W} \times \mathcal{T}$ that is independent of X and has intensity function $\rho(\boldsymbol{w}, t)$. Applying the Campbell-Mecke theorem (2.6) to D, one has

$$\mathbb{E}\left[\sum_{\boldsymbol{x} \in D} \frac{1}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})} \nabla \lambda(\boldsymbol{x}; \boldsymbol{\theta})\right] = \int_{\mathcal{W}} \int_{\mathcal{T}} \frac{\rho(\boldsymbol{w}, t) \nabla \lambda(\boldsymbol{w}, t; \boldsymbol{\theta})}{\lambda(\boldsymbol{w}, t; \boldsymbol{\theta}) + \rho(\boldsymbol{w}, t)} \mathrm{d}\boldsymbol{w} \mathrm{d}t. \tag{2.8}$$

Having obtained unbiased estimators for both left and right hand side of equation (2.6) with $\boldsymbol{f}(\boldsymbol{w}, t)$ given by (2.7), one can plug them in to obtain the estimating equation. It solves

$$\begin{aligned} \boldsymbol{s}(X, D; \boldsymbol{\theta}) = &\sum_{\boldsymbol{x} \in X} \frac{\rho(\boldsymbol{x})/\lambda(\boldsymbol{x}; \boldsymbol{\theta})}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})} \nabla \lambda(\boldsymbol{x}; \boldsymbol{\theta}) \\ &- \sum_{\boldsymbol{x} \in D} \frac{1}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})} \nabla \lambda(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{0} \end{aligned} \tag{2.9}$$

over the parameter $\boldsymbol{\theta}$. Note that the subscript $\boldsymbol{x}$ in the two sums runs through the points of X and D, respectively. It is interesting to observe that (2.9) is precisely the derivative of a logistic log-likelihood function

$$L(X, D; \boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in X} \log \left[\frac{\lambda(\boldsymbol{x}; \boldsymbol{\theta})}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})}\right] + \sum_{\boldsymbol{x} \in D} \log \left[\frac{\rho(\boldsymbol{x})}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})}\right]. \tag{2.10}$$

In this way, the fitting of a Poisson point process model can be transformed into a maximum likelihood estimation problem of logistic regression. It can be easily implemented based on the R-package *stats* (Venables & Ripley, 2002) and treated

as the fitting of generalized linear models with a *logit* link function. Since the log-likelihood function is concave, the existence and uniqueness of parameter $\boldsymbol{\theta}$ that maximizes the log-likelihood are guaranteed under certain conditions (Silvapulle, 1981).

It is worth noting that the point process model itself is not a logistic model, only the estimation takes the form of logistic regression.

2.4.4 Modelling Chimney Fire Data

As introduced above, we use logistic regression estimation to fit the intensity functions of our four house types separately. Consider (2.4) for fixed k and write $\boldsymbol{\theta}_k$ as above for the vector of coefficients in the harmonic and polynomial functions. Also, consider the chimney fires occurring in houses of type k as a point process X_k with intensity function $\lambda_k(\boldsymbol{w}, t; \boldsymbol{\theta}_k)$. Denote the corresponding dummy point process by D_k with intensity function $\rho_k(\boldsymbol{w}, t)$. Assuming that we have p parameters in $\boldsymbol{\theta}_k$, the estimating equations then read

$$s_i(X_k, D_k; \boldsymbol{\theta}_k) = \sum_{\boldsymbol{x} \in X_k} \frac{\rho_k(\boldsymbol{x}) z_i(\boldsymbol{x})}{\lambda_k(\boldsymbol{x}; \boldsymbol{\theta}_k) + \rho_k(\boldsymbol{x})} - \sum_{\boldsymbol{x} \in D_k} \frac{\lambda_k(\boldsymbol{x}; \boldsymbol{\theta}_k) z_i(\boldsymbol{x})}{\lambda_k(\boldsymbol{x}; \boldsymbol{\theta}_k) + \rho_k(\boldsymbol{x})} = 0 \quad (2.11)$$

with $p = 1, \ldots, m$, where $z_i(\boldsymbol{x})$ denotes the i-th temporal covariate (i.e., the harmonic components, wind speed and wind chill terms with different orders and the interaction terms of wind chill and wind speed in (2.4)) at point $\boldsymbol{x}$.

To estimate the parameters in $\boldsymbol{\theta}_k$, we perform a three-step implementation. First, we specify the intensity function $\rho_k(\boldsymbol{w}, t)$ of the dummy point process D_k. Note that the role of the dummy point process is to approximate the integral in the Campbell–Mecke theorem (2.6) by (2.8). In order to efficiently estimate the parameters, we need to put more dummy points in the regions and times with a relatively high number of fire occurrences so that the approximation of the integral in those parts can be more accurate. Considering that, spatially, most chimney fires occur in urban areas, and that temporally, there are more chimney fires in winter than in summer, specifically, we use the density map of houses of the given type k (cf., Figure 2.7) to distinguish urban areas from rural areas and employ a sine function to assign higher $\rho_k(\boldsymbol{w}, t)$ to cold seasons to comply with the observations in Figure 2.2. Thus, the intensity function $\rho_k(\boldsymbol{w}, t)$ of D_k is given by

$$\rho_k(\boldsymbol{w}, t) = r_k h_k(\boldsymbol{w}) \left\{ 0.5 + 0.25 \left[\sin\left(\frac{2\pi}{365} t + \frac{\pi}{2} \right) + 1 \right] \right\}, \quad (2.12)$$

where r_k is a multiplication factor used to ensure that $\rho_k(\boldsymbol{w}, t)$ is at least four times $\lambda_k(\boldsymbol{w}, t; \boldsymbol{\theta}_k)$ as suggested in Baddeley et al. (2014) and $h_k(\boldsymbol{w})$ is the density of houses of type k at location $\boldsymbol{w}$. Specifically, we set r_k's to 60, 20, 20 and 8 for the four house types, respectively. Second, we generate a realization from the dummy point process D_k using the R-package *spatstat* (Baddeley, Rubak & Turner, 2015). Thirdly, based on the observation of X_k and the realization of

the dummy point process D_k, we use the R-package *stats* (Venables & Ripley, 2002) to estimate model parameters $\boldsymbol{\theta}_k$. In addition, in order to determine the optimal function orders, $o_{k,1}$, $o_{k,2}$, $o_{k,3}$, $o_{k,4}$, in (2.4) simultaneously, we apply a grid search to select the combination that yields the smallest Akaike information criterion (Choiruddin, Coeurjolly & Waagepetersen, 2020) given a set of ranges – $o_{k,1}$: 1–4, $o_{k,2}$: 1–5, $o_{k,3}$: 1–5, $o_{k,4}$: 1–5 – for them. Considering that wind chill already contains some information of wind speed, we also perform likelihood ratio tests over the best models with and without wind speed and the interaction terms.

We separate the fire data into two sets: the data on the period 2004–2019 and the data on the year 2020, for fitting and predicting purposes, respectively. Performing the model selection introduced above on the data over 2004–2019, we obtain the models for each of the four house types. We find that wind chill alone is mostly sufficient to obtain accurate risk predictions, except for house type 4, where including the interaction terms provides a better fitting performance. Finally, the temporal intensity functions obtained for the four house types read as follows:

$$\begin{aligned}\phi_1(t;\boldsymbol{\theta}_1) = \exp\Bigg[&\theta_{1,1} + \theta_{1,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{1,3}\sin\left(\frac{2\pi}{365}t\right)\\ &+ \theta_{1,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{1,5}\sin\left(\frac{4\pi}{365}t\right) + \theta_{1,6}\cos\left(\frac{6\pi}{365}t\right)\\ &+ \theta_{1,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{1,8}\cos\left(\frac{8\pi}{365}t\right) + \theta_{1,9}\sin\left(\frac{8\pi}{365}t\right)\\ &+ \theta_{1,10}V_{\tau,3}(t) + \theta_{1,11}V_{\tau,3}^2(t)\Bigg],\end{aligned} \tag{2.13}$$

$$\begin{aligned}\phi_2(t;\boldsymbol{\theta}_2) = \exp\Bigg[&\theta_{2,1} + \theta_{2,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{2,3}\sin\left(\frac{2\pi}{365}t\right)\\ &+ \theta_{2,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{2,5}\sin\left(\frac{4\pi}{365}t\right) + \theta_{2,6}\cos\left(\frac{6\pi}{365}t\right)\\ &+ \theta_{2,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{2,8}V_{\tau,3}(t) + \theta_{2,9}V_{\tau,3}^2(t)\\ &+ \theta_{2,10}V_{\tau,3}^3(t) + \theta_{2,11}V_{\tau,3}^4(t)\Bigg],\end{aligned} \tag{2.14}$$

$$\begin{aligned}\phi_3(t;\boldsymbol{\theta}_3) = \exp\Bigg[&\theta_{3,1} + \theta_{3,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{3,3}\sin\left(\frac{2\pi}{365}t\right)\\ &+ \theta_{3,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{3,5}\sin\left(\frac{4\pi}{365}t\right) + \theta_{3,6}\cos\left(\frac{6\pi}{365}t\right)\\ &+ \theta_{3,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{3,8}V_{\tau,3}(t)\Bigg],\end{aligned} \tag{2.15}$$

$$\begin{aligned}\phi_4(t;\boldsymbol{\theta}_4) = \exp\Bigg[&\theta_{4,1} + \theta_{4,2}\cos\left(\frac{2\pi}{365}t\right) + \theta_{4,3}\sin\left(\frac{2\pi}{365}t\right)\\ &+ \theta_{4,4}\cos\left(\frac{4\pi}{365}t\right) + \theta_{4,5}\sin\left(\frac{4\pi}{365}t\right) + \theta_{4,6}\cos\left(\frac{6\pi}{365}t\right)\\ &+ \theta_{4,7}\sin\left(\frac{6\pi}{365}t\right) + \theta_{4,8}\cos\left(\frac{8\pi}{365}t\right) + \theta_{4,9}\sin\left(\frac{8\pi}{365}t\right)\\ &+ \theta_{4,10}V_{\tau,3}(t) + \theta_{4,11}V_{\tau,3}^2(t)\\ &+ \theta_{4,12}V_{\tau,3}^3(t) + \theta_{4,13}V_{\tau,1}(t)V_{\tau,3}(t)\Bigg].\end{aligned} \tag{2.16}$$

We provide the estimates of all model parameters (i.e., $\boldsymbol{\theta}$'s) in Table 2.2. Moreover, we plot the spatial and temporal predictions and the actual realizations for the year 2020 in Figure 2.8(top). The spatial realization is smoothed by Gaussian kernels with a standard deviation of 1000 metres.

Table 2.2 Parameter estimates for the Poisson point process model defined by the intensity functions of (2.13)–(2.16) and their 95% confidence intervals (CI). The 'e' denotes a base of 10.

Parameter	Estimate (CI)	Parameter	Estimate (CI)
$\theta_{1,1}$	-1.22e1(±3.82e-1)	$\theta_{1,2}$	-2.78e-1(± 4.81e-1)
$\theta_{1,3}$	-1.44e-1(± 2.84e-1)	$\theta_{1,4}$	-1.63e-1(± 2.64e-1)
$\theta_{1,5}$	7.99e-2(± 3.04e-1)	$\theta_{1,6}$	-9.21e-3(± 2.49e-1)
$\theta_{1,7}$	-2.33e-2(± 2.62e-1)	$\theta_{1,8}$	3.05e-1(± 2.19e-1)
$\theta_{1,9}$	1.59e-2(± 2.20e-1)	$\theta_{1,10}$	-7.42e-2(± 3.56e-2)
$\theta_{1,11}$	-6.12e-3(± 3.31e-3)	$\theta_{2,1}$	-1.30e1(± 2.31e-1)
$\theta_{2,2}$	6.40e-2(± 2.74e-1)	$\theta_{2,3}$	1.42e-2(± 1.56e-1)
$\theta_{2,4}$	-3.14e-2(± 1.64e-1)	$\theta_{2,5}$	1.69e-1(± 1.60e-1)
$\theta_{2,6}$	1.48e-1(± 1.24e-1)	$\theta_{2,7}$	-4.54e-2(± 1.27e-1)
$\theta_{2,8}$	-6.81e-2(± 2.64e-2)	$\theta_{2,9}$	2.33e-3(± 3.65e-3)
$\theta_{2,10}$	-8.50e-6(± 2.47e-4)	$\theta_{2,11}$	-2.76e-5(± 1.93e-5)
$\theta_{3,1}$	-1.43e1(± 9.48e-1)	$\theta_{3,2}$	1.57e0(± 1.60e0)
$\theta_{3,3}$	2.25e-1(± 6.00e-1)	$\theta_{3,4}$	-1.19e0(± 1.09e0)
$\theta_{3,5}$	-3.19e-1(± 7.20e-1)	$\theta_{3,6}$	6.32e-1(± 6.06e-1)
$\theta_{3,7}$	1.65e-1(± 5.39e-1)	$\theta_{3,8}$	-1.12e-1(± 5.25e-2)
$\theta_{4,1}$	-1.39e1(± 2.41e-1)	$\theta_{4,2}$	1.49e-1(± 3.05e-1)
$\theta_{4,3}$	6.11e-2(± 1.76e-1)	$\theta_{4,4}$	-2.06e-1(± 1.83e-1)
$\theta_{4,5}$	9.00e-2(± 1.95e-1)	$\theta_{4,6}$	2.77e-2(± 1.61e-1)
$\theta_{4,7}$	7.11e-2(± 1.72e-1)	$\theta_{4,8}$	1.82e-1(± 1.36e-1)
$\theta_{4,9}$	-4.84e-2(± 1.38e-1)	$\theta_{4,10}$	-9.38e-2(± 4.06e-2)
$\theta_{4,11}$	-7.92e-4(± 1.99e-3)	$\theta_{4,12}$	-2.35e-4(± 1.62e-4)
$\theta_{4,13}$	2.99e-3(± 2.10e-3)		

Overall, both spatial and temporal predictions capture the correct trends of fire occurrences. Specifically, in the spatial domain, our model learns the urbanity features of the Twente region which are highly relevant for chimney fires. Since the spatial prediction depends on the distribution of houses of different types which can only be accessed as actual data, our model displays similar spatial patterns for different years. In the temporal domain, our model captures the periodic pattern on an annual basis and adds weather variable dependent

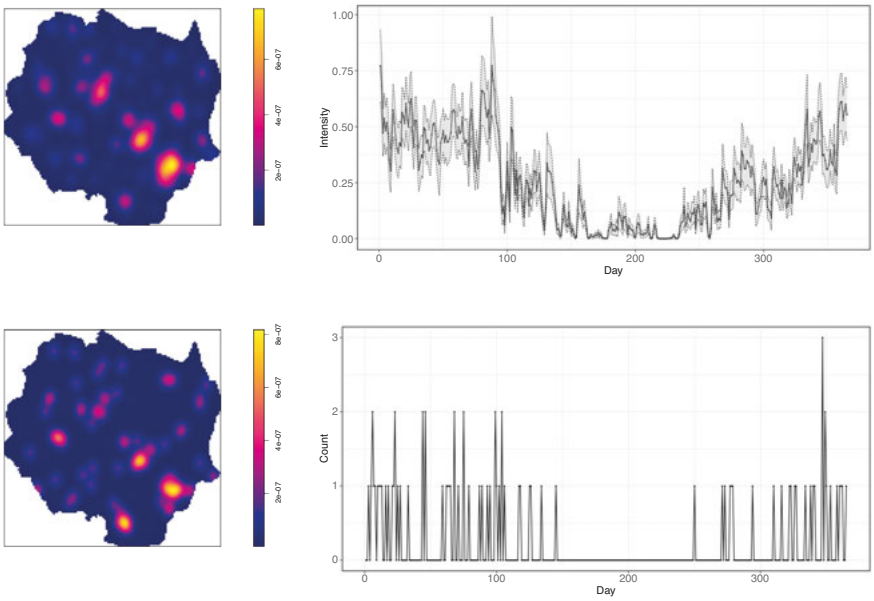


Figure 2.8 Spatial and temporal predictions (top left, top right) based on our point process model and actual realizations (bottom left, bottom right) for the year 2020. Shadows in the temporal predictions of the point process model bound the 95% confidence intervals. The unit in spatial plots is metre^{-2}; the unit in temporal plots is day^{-1}.

information, which explains the noisier aspect of Figure 2.8(top right) compared to Figure 2.8(top left). A comparison to the areal unit model and some ρ-tuning experiments will be discussed in Section 2.4.6 and Section 2.6.2. In summary, given appropriate building information, our prediction model can detect areas with higher risks of chimney fires in Twente and, additionally, estimate the risk for specific days based on weather forecast.

2.4.5 Confidence Interval

For risk prediction problems, it is important to quantify the uncertainty of the estimates. In the literature, Baddeley et al. (2014) demonstrated asymptotic normality for their logistic regression estimators for stationary Gibbs point processes in space when the domain increases. Recently, Choiruddin, Coeurjolly & Letué (2018) proved asymptotic normality also in an increasing-domain setting for spatial point process models that involve explanatory variables. For chimney fires, the increasing-domain setting is unrealistic as fire patterns can vary considerably from region to region. In Chapter 3, we will study strong consistency and asymptotic normality for the logistic regression estimators we use above for

spatio-temporal point processes in an infill regime.

Specifically, under appropriate conditions, we prove that the maximizer $\widehat{\boldsymbol{\theta}}$ of (2.11), under the true value $\boldsymbol{\theta_0}$, is approximately normally distributed with mean $\boldsymbol{\theta_0}$ and covariance matrix $\boldsymbol{G}$ given by the inverse of the Godambe matrix (Godambe & Heyde, 2010). A plug-in estimator for $\boldsymbol{G}$ is

$$\widehat{\boldsymbol{G}} = \left[\int_{\mathcal{W}} \int_{\mathcal{T}} \frac{\lambda(\boldsymbol{w}, t; \widehat{\boldsymbol{\theta}}) \rho(\boldsymbol{w}, t)}{\lambda(\boldsymbol{w}, t; \widehat{\boldsymbol{\theta}}) + \rho(\boldsymbol{w}, t)} z_i(\boldsymbol{w}, t) z_j(\boldsymbol{w}, t) \mathrm{d}\boldsymbol{w} \mathrm{d}t \right]_{i,j=1}^{p}, \tag{2.17}$$

assuming that we have p parameters in $\boldsymbol{\theta}$. Approximate confidence intervals for model parameters are then readily obtained.

For chimney fire data modelling, the approximate numerical 95% confidence intervals of model parameters are listed in Table 2.2. The Delta method (Ver Hoef, 2012) can again be used to calculate approximate confidence intervals for predicted fire intensities. We visualize the temporal risk predictions for the year 2020 in Figure 2.8(top right) as an example.

2.4.6 Comparison to an Areal Unit Model

As a pilot study, we propose an areal unit fire prediction model using the same selected variables as described in Section 2.3. In contrast to the data preparation procedure introduced in Section 2.2, the fire data and environmental variables for this model are aggregated at the level of $500m \times 500m$ area boxes and daily time intervals. As a result, it has a different notation:

- $N_{w,t}$: the number of chimney fires occurring in box w on day t,
- $V_{\sigma,i,w}$: the value of the i-th spatial variable in box w,
- $V_{\tau,i,t}$: the value of the i-th temporal variable on day t.

Here, σ and τ again refer to spatial and temporal variables, cf. Table 2.1. We also divide the houses into four types based on age and on whether or not they are freestanding, leading to the following Poisson model

$$N_{w,t} = \sum_k N_{w,t}^k \sim \text{Poisson}(\sum_k h_w^k \lambda_t^k), \tag{2.18}$$

with $k = 1, \ldots, 4$, where h_w^k represents the number of house of type k in box w and λ_t^k is the fire risk intensity for house of type k on day t. The intensity function λ_t^k is defined as

$$\begin{aligned} \lambda_t^k = &\exp\big[\text{Harmonic}(t, o_h^k) + \text{Polynom}(V_{\tau,1,t}, o_{p1}^k) + \\ &\text{Polynom}(V_{\tau,3,t}, o_{p2}^k) + \text{Polynom}(V_{\tau,1,t} V_{\tau,3,t}, o_{p3}^k)\big]. \end{aligned} \tag{2.19}$$

We fit this model by maximum likelihood estimation and select the function orders $o_h^k, o_{p1}^k, o_{p2}^k, o_{p3}^k$ based on the Akaike information criterion. To compare the areal unit model to the point process model, we again estimate parameters in the former using the data on the period 2004–2019 and test its performance

for the year 2020. We establish theoretical confidence intervals for parameter estimates and predicted fire intensities based on classical theories: the Fisher information for asymptotic normality for generalized linear models (Fahrmeir & Kaufmann, 1985; McCullagh & Nelder, 2019) and the Delta method (Ver Hoef, 2012).

The spatial and temporal predictions based on the areal unit model for the year 2020 are plotted in Figure 2.9. Since the spatial predictions from the areal unit model are computed for $500m \times 500m$ area boxes, we average them by the volume of these boxes to comply with the scaling level of the point process model. Generally, comparing Figure 2.9 to Figure 2.8, we see that both areal unit model and point process model capture similar spatial and temporal patterns from the fire data. However, the spatial predictions from the point process model by construction are smoother than those from the areal unit model, as we apply kernel smoothing in the former to obtain the density of a house type at a location, which helps establish spatially continuous predictions. The temporal predictions from the two models are almost identical for every day, because we assume that temporal variables at different times of a day remain invariant and the spatial aggregation is integrated out. Specifically, in terms of the total chimney fire risk predicted for the year 2020, the areal unit model predicts 92.86 and the point process model predicts 92.95, which are very close to each other. The observed actual count of 81 lies in the 95% confidence intervals of both models. However, the areal unit model overestimates many more fire risks for certain area boxes of big city centres (e.g., Enschede) and underestimates for rural area boxes.

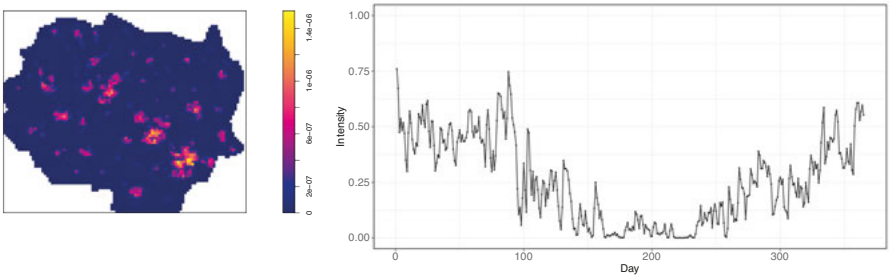


Figure 2.9 Spatial (left) and temporal (right) predictions based on the areal unit model for the year 2020. The unit in the spatial plot is metre^{-2}; the unit in the temporal plot is day^{-1}.

To conclude, the point process model generates continuous hazard maps of chimney fires, which avoids the classic modifiable area unit data problem and enables to access fire risks at exact locations in Twente, thus is more user-friendly for fire services. The predictions do not depend on the scale of the area boxes, while only the smoothing bandwidth is inevitably inherited. Moreover, from a computational perspective, the logistic regression estimation approach used to fit the point process model is more efficient and captures more accurate data

patterns.

2.5 Model Validation

In this section, we validate the independence assumption underlying our Poisson point process model by analyzing the second-order properties of the chimney fire data. We also validate the first-order structure of our model by visualizing the residuals.

2.5.1 Second-order Analysis

To assess the validity of the Poisson assumption, we perform a second-order analysis on the fire data using two summary statistics: the pair correlation function and the K-function (cf., Section 1.1). The pair correlation function can help detect inter-point interactions for specific distances, whereas the K-function is a cumulative statistic that can be used to test the existence of point interactions in data.

The pair correlation function

For visualization purposes, we test pair correlation functions for the projections in space and time, separately. In the spatial case, under appropriate stationarity and isotropy assumptions (Baddeley, Møller & Waagepetersen, 2000), the pair correlation function is a function of the spatial distance between two points, $r = ||\boldsymbol{x} - \boldsymbol{y}||$, and can be estimated by

$$\widehat{g}(r) = \frac{1}{\pi r} \sum_{\boldsymbol{x} \in X} \sum_{\boldsymbol{y} \in X, \boldsymbol{y} \neq \boldsymbol{x}} \frac{k_b(r - ||\boldsymbol{x} - \boldsymbol{y}||)}{\lambda(\boldsymbol{x})\lambda(\boldsymbol{y})|\mathcal{W} \cap \mathcal{W}_{\boldsymbol{x}-\boldsymbol{y}}|}, \tag{2.20}$$

where $k_b(\cdot)$ is a one-dimensional smoothing kernel with bandwidth b and $1/|\mathcal{W} \cap \mathcal{W}_{\boldsymbol{x}-\boldsymbol{y}}|$ is the spatial edge correction factor (Ohser & Stoyan, 1981). For temporal use, $||\boldsymbol{x} - \boldsymbol{y}||$ becomes the time interval between two points, and one needs to change πr to one and $1/|\mathcal{W} \cap \mathcal{W}_{\boldsymbol{x}-\boldsymbol{y}}|$ to $1/|\mathcal{T} \cap \mathcal{T}_{\boldsymbol{x}-\boldsymbol{y}}|$, accordingly. If there is no interaction, the pair correlation function is one. At short distances, a value of $\widehat{g}(r)$ larger than one suggests clustering in the sense that points are more likely to occur from one another, whereas a value smaller than one suggests inhibition, that is, points tend to mutually avoid each other. If the intensity function $\lambda(\boldsymbol{x})$ is unknown, a plug-in estimator can be used.

In practice, we employ (2.20) with Gaussian smoothing kernels as k_b and plug in the fitted intensity functions based on the data from all 17 years. We set the possible interaction intervals for the spatial and temporal domain to 10000 metres and 100 days, respectively, and the smoothing bandwidths to 500 metres and 10 days. The estimated pair correlation functions are plotted in Figure 2.10. Overall, both the spatial and temporal results show a curve that converges to one. At smaller distances, there is some attraction between the points spatially.

Temporally, possibly because of the precision of the observations (daily basis), there is some repulsion and a weak peak at an interval of about twenty days. The latter might also be due to underestimates of the intensity function on a specific pair of days, cf. (2.20).

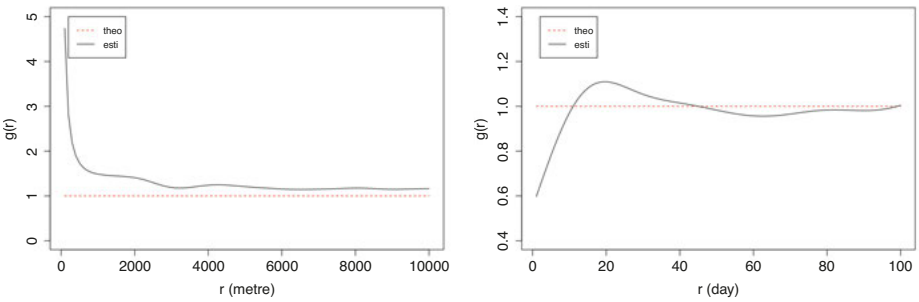


Figure 2.10 Estimated (solid black line) and theoretical (dashed red line) pair correlation functions for the spatial and temporal projections (cf., Figure 2.1(right) and 2.2) of the chimney fire incidents during 2004–2020 using the intensity functions fitted by our Poisson point process model.

The K-function

Since the pair correlation function shows some point interactions at small distances and time intervals, we perform a joint space-time K-function test on the fire data. Under appropriate weak stationarity assumptions (Gabriel & Diggle, 2009), a point-wise unbiased estimator of the inhomogeneous K-function is

$$\widehat{K}(r,v) = \sum_{\boldsymbol{x}\in X} \sum_{\boldsymbol{y}\in X, \boldsymbol{y}\neq\boldsymbol{x}} \frac{1(||\boldsymbol{w}(\boldsymbol{x}) - \boldsymbol{w}(\boldsymbol{y})|| \leq r, |t(\boldsymbol{x}) - t(\boldsymbol{y})| \leq v)}{\lambda(\boldsymbol{x})\lambda(\boldsymbol{y})|(\mathcal{W}\times\mathcal{T}) \cap (\mathcal{W}\times\mathcal{T})_{\boldsymbol{x}-\boldsymbol{y}}|}, \tag{2.21}$$

where $\boldsymbol{w}(\boldsymbol{x}), \boldsymbol{w}(\boldsymbol{x})$ and $t(\boldsymbol{x}), t(\boldsymbol{y})$ represent the spatial and temporal coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, $1(\cdot)$ is the indicator function and the other notations are as in the previous section. In the absence of interaction, $K(r,v) = 2\pi r^2 v$. Moreover, when the intensity function $\lambda(\boldsymbol{x})$ is unknown, a plug-in estimator can be used.

We compute the inhomogeneous K-function over the chimney fire data from all 17 years. To obtain an elegant view, we set the testing sequences of spatial and temporal distances to 10000 metres and 100 days and divide both of them into 100 step size pairs of 100 metres and 1 day. We calculate the K-function using the fitted intensity functions by our Poisson point process model to estimate λ in (2.21) and plot the result in Figure 2.11(left). As for the pair correlation function, we find some evidence of clustering. To test whether the interaction is statistically significant, we implement a Monte Carlo test (Baddeley, Rubak & Turner, 2015). We generate 99 realizations of a Poisson point process with the

fitted intensities (i.e., by our Poisson point process model) and compute their K-functions and that of the actual chimney fire data but now using kernel estimators to estimate the intensity function λ in (2.21). We plot the local envelopes in Figure 2.11(right). As the empirical K-function of the actual chimney fire data lies completely within the envelope, there is no reason to look beyond a Poisson model.

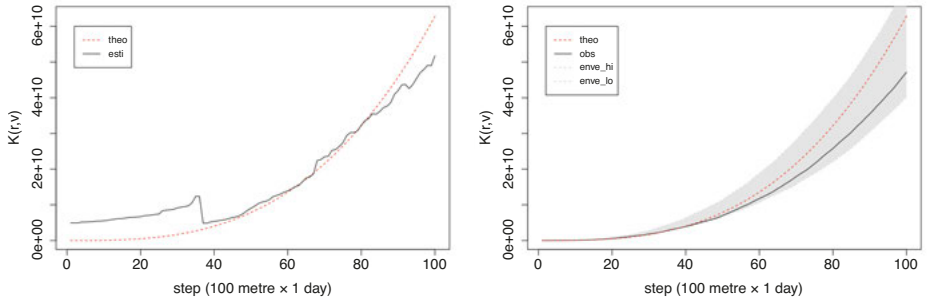


Figure 2.11 Estimated (solid black line) and theoretical (dashed red line) inhomogeneous K-functions for the space-time pattern of the chimney fire incidents during 2004–2020 using the intensity functions fitted by our Poisson point process model (left). Estimated (solid black line) and theoretical (dashed red line) inhomogeneous K-functions for the chimney fire data and the envelope of the estimated inhomogeneous K-functions (shadows) for the 99 Monte Carlo simulations of a Poisson point process with the fitted intensities, both using the intensity functions fitted by kernel estimators (right).

2.5.2 Residual Analysis

In order to further verify the validity of our point process model, we perform a residual analysis to check the remaining structure after the modelling of a Poisson point process with specific forms defined in (2.3), (2.4) and (2.13)–(2.16). We fit the model using the data from all 17 years and compute spatial and temporal residuals separately. As suggested in Baddeley et al. (2005), the spatial residuals are defined as the difference between the smoothed fitted fire risk and the smoothed actual realizations using Gaussian kernels with a standard deviation of 1000 metres. Temporally, we show the residuals of the difference between the predicted and actual monthly counts.

The results of the residual analysis are plotted in Figure 2.12. In the spatial domain, no specific pattern is shown, except that our Poisson point process model still overestimates for big city centres (e.g., Enschede and Hengelo) whereas it underestimates a bit for small city centres. This could be expected, as our model depends only on the density of houses of various types and we observe a saturation pattern of chimney fire occurrences when the density of a house type increases (cf., Figure 2.6). A possible solution for that is to apply a piece-wise function instead

of the linear function to model the relation between chimney fire occurrences and the density of a house type. In the temporal domain, no particular pattern is apparent.

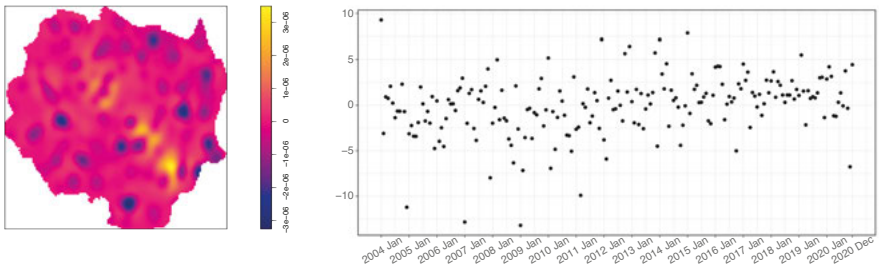


Figure 2.12 Spatial (left) and temporal (right) residuals based on the chimney fire data during 2004–2020. The unit in the spatial plot is metre^{-2}; the unit in the temporal plot is month^{-1}.

2.6 Discussion

In this section, we discuss the data-driven modelling procedure designed for chimney fire prediction, which can be generalized to similar fire patterns. We also investigate the influence of tuning the dummy intensity ρ to important regions and times in logistic regression estimation on model fitting.

2.6.1 Modelling Procedure

The data-driven modelling procedure is two-step. In the first step, we combine machine learning and statistics, where the former is utilized to help select the most important environmental variables and the latter appropriately approximates and explains the relations between explanatory variables and fire incidence. We employ parametric functions to model such relations and fit them using logistic regression estimation, so that the influence of a variable on chimney fires can be analyzed by its coefficient in the function model. This combination offers significant advantages. From the machine learning side: (i) compared to naive statistics used for the selection of explanatory variables, random forests can detect both linear and non-linear correlation between a candidate variable and chimney fire occurrences; (ii) random forests consider the dependence among candidate variables nonparametrically; (iii) through applying the conditional permutation importance techniques instead of the traditional ones, the bias towards correlated variables is suppressed; (iv) machine learning algorithms can handle the selection with a large number of putative variables while statistical methods often behave

unstable or even fail in such cases. From the statistics side: (i) classic statistical methods for regression modelling with specific mathematical expressions are more interpretable and based on natural assumptions; (ii) statistical methods also allow for significance testing for detailed analyses and uncertainty quantification. However, due to the fact that the environmental variables relevant to our chimney fire data are not available for every point in space and time, we perform variable selection and model fitting at different levels of data aggregation, which has the disadvantage that the selected explanatory variables might not be consistent. Here, we leave it as an important direction in future work.

In the second step, we use summary statistics, the pair correlation function and the K-function, to detect point interactions. We also perform a residual analysis to validate the model structure. Although, in our study, a Poisson point process model is sufficient for chimney fire prediction, a hierarchical model (Banerjee, Carlin & Gelfand, 2014), where random effects are incorporated to deal with latent variables, might be useful for other fire types.

2.6.2 ρ Tuning in Logistic Regression Estimation

Recall that the role of the dummy point process in logistic regression estimation is to estimate the integral in (2.6). In Baddeley et al. (2014), a rule of thumb was proposed that the intensity function ρ of the dummy point process should be at least four times the intensity function λ of the point process to be estimated; additionally, a data-driven selection of ρ was suggested.

In Section 2.4.4, we tuned ρ to concentrate on the urban areas and cold seasons by (2.12) in order to efficiently obtain more precise estimates for model parameters provided a limited number of dummy points. Here, we perform two experiments to verify the influence of this ρ tuning operation in the spatial and temporal domains separately. Specifically, we fit our fire prediction model on the data from all 17 years using logistic regression estimation with and without tuning the dummy intensity ρ and plot the difference of the fitted values. For comparison, we set the expectations of the number of dummy points as equal in both cases. In addition, since the realizations of the dummy point process generated in different runs contain randomness, we perform each experiment 60 times and plot the averaged spatial and temporal differences to reduce the randomness.

The results show that, under the premise of the same amount of calculation, concentrating on important regions and times enables the model to obtain more accurate patterns of chimney fires. Spatially, according to Figure 2.13(left), the bias caused by the saturation effect shown in Figure 2.6 is reduced. Temporally, plotted in Figure 2.13(right), the fitted fire intensities increase in March and April, which are known as the weather tipping seasons and are found more likely to catch chimney fires in School (2018). To summarize, tuning the dummy intensity ρ to regions and times that are salient for fire risk is helpful to estimate the risk more efficiently and capture important underlying patterns.

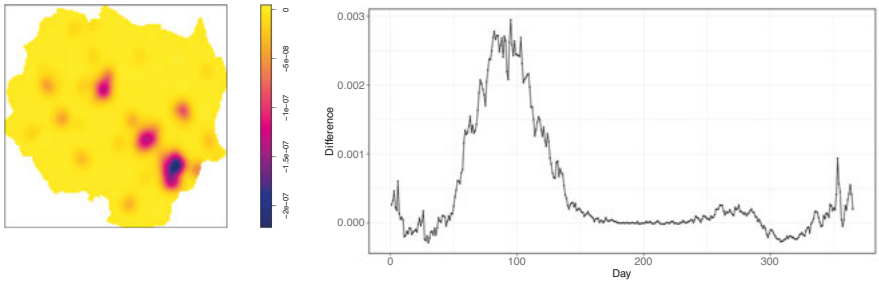


Figure 2.13 Spatial (left) and temporal (right) difference of the fitted intensities using tuned ρ compared to uniform ρ with an equal expected number of dummy points in logistic regression estimation (averaged on 60 random runs). The unit in the spatial plot is metre^{-2}; the unit in the temporal plot is day^{-1}.

2.7 Extended Applications

To demonstrate the generalizability of our proposed approach, this section presents applications of our approach to other types of fires and chimney fire patterns in other geographical regions.

2.7.1 Kitchen Fire Prediction in Twente

The first application is concerned with kitchen fire prediction in the same region, Twente, whose point pattern is plotted in Van Leeuwen (2022). The fire incidents, along with the relevant environmental variables, were collected in the same manner as introduced in Section 2.2. Similar to chimney fires, kitchen fires show pronounced spatial heterogeneity, with more incidents occurring in densely populated areas (cf., Figure 2.1(right)). Unlike chimney fires, they do not exhibit a clear temporal pattern (cf., Figure 2.2), thus the risk prediction was restricted to the spatial domain. In addition to the 22 spatial variables listed in Table 2.1, the experts from the Twente Fire Brigade suggested five additional putative covariates: the percentage of residents in the first and last 20% of the income distribution as well as the average property value and the average gas and electricity consumptions in the neighbourhood.

Van Leeuwen (2022) adapted the areal unit method in Section 2.4.6 to predict the kitchen fire risk. She conducted the same variable selection analysis with random forests and permutation importance techniques. Her results revealed multiple variables with high importance scores, contrasting with the straightforward results obtained for chimney fires in Section 2.3. While key variables were identified, formulating an interpretable model structure as in Section 2.4.2 proved more challenging. Instead of replicating the areal unit model for chimney fires, Van Leeuwen (2022) constructed a stepwise Poisson regression model, selecting the best-fitting variable components based on the Akaike information

criterion. Although the final predictions were plausible, this case study concludes that our proposed modelling procedure may not always produce an interpretable and well-structured statistical model for fire risk prediction, as variable selection and model formulation are performed separately.

2.7.2 Chimney Fire Prediction in IJsselland

In the second application, Van Kasteren (2024) tested our approach for predicting the chimney fire risk in IJsselland, a neighbouring region of Twente. Although geographically close, IJsselland exhibits distinct characteristics: (i) it has a higher concentration of historical buildings, e.g. churches and museums; (ii) the cities are more centralized and densely populated. Again, he collected the fire incidents and environmental variables as in Section 2.2, with the exception that the temporal variable of visibility was unavailable. The spatial and temporal projections of chimney fires in IJsselland presented in Van Kasteren (2024) display similar patterns to those in Twente.

Different from the kitchen fire prediction study, Van Kasteren (2024) considered a point process model to predict chimney fires in IJsselland. His variable selection results were similar to those in Section 2.3, but with specific features unique to IJsselland. Specifically, he identified old, freestanding houses and wind chill as the most significant variables. However, unlike in Twente, the number of buildings constructed before 1920 also emerged as an important factor, which may be due to the presence of historical buildings in IJsselland. Following our work in this chapter, Van Kasteren (2024) divided the buildings into six types, developed a corresponding Poisson point process model as in Section 2.4.2, applied the logistic regression estimation for model fitting, and achieved promising predictions. This case study shows that our proposed modelling procedure can be effectively applied to model similar fire patterns in geographical regions with different underlying characteristics.

2.8 Summary

In this chapter, we proposed a data-driven modelling procedure for chimney fire risk prediction, which can be generalized to studies on other fire data that involves covariate-based phenomenon. First, we applied random forests and permutation importance techniques to select the explanatory variables nonparametrically from a large number of environmental variables. From a practical perspective, our results indicate that pre-war detached or semi-detached houses run a higher risk of chimney fires and therefore public awareness campaigns should preferentially target the owners of such houses. Second, based on the observed relations between the selected variables and chimney fire occurrences, we defined a Poisson point process model to learn the observed spatio-temporal point pattern in order to predict fire risk. We applied logistic regression estimation for model fitting and provided infill asymptotic confidence intervals, the construction of which will

be studied in Chapter 3. Additionally, to validate our model assumptions, we performed a second-order analysis and a residual study. Last but not least, we reviewed our modelling procedure and showed the advantages of the tuned intensity function of the dummy point process in logistic regression estimation.

For future work, as presented in Section 2.7, the first direction would be to extend the data-driven modelling procedure to a broader range of fire types to assess its applicability. Additionally, in this chapter, we performed our nonparametric variable selection on areal unit data while fitted our statistical model on point process data, which may lead to inconsistencies due to different levels of data aggregation. Hence, it would also be valuable to develop a fully integrated data-driven approach for point process modelling that combines both variable selection and intensity estimation within a single framework.

2.9 Appendix

Results for additional variable importance experiments

In the selection of explanatory variables, to test the influence of different hyperparameter settings of random forests (i.e., the number of randomly sampled input variables at each tree node, mtry, and the random seed) on variable importance, we particularly perform experiments that construct random forests with multiple configurations. Moreover, to assess the influence of small variations in weather among different parts of the Twente region on variable importance, we perform another analysis on the temporal data from two neighbouring weather stations.

The variable importance results for testing the influence of different hyperparameter settings of random forests on variable importance are plotted in Figures 2.14 and 2.15. It shows that both mtry and random seed settings may influence the detailed importance scores (the increase of the prediction error) of some variables, but they will not influence the detection of the most significant variables. The variable importance results for assessing the influence of small variations in weather among different parts of the Twente region on variable importance are plotted in Figure 2.16. The significant variables do not change, which indicates that those small variations can be ignored.

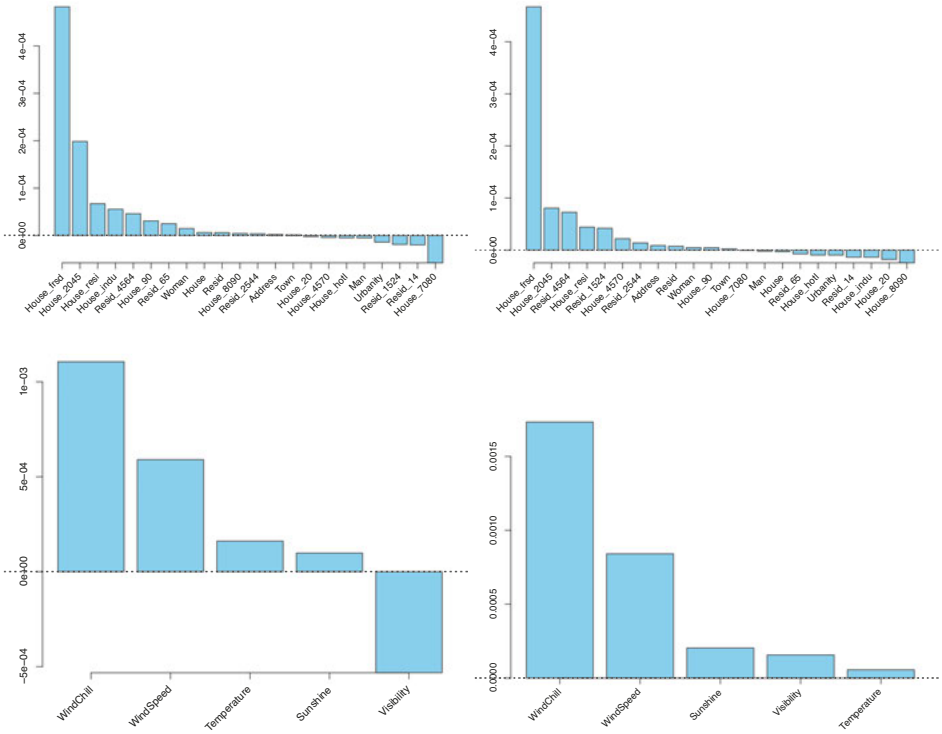


Figure 2.14 Tests on the influence of the hyperparameter – mtry – in the selection of explanatory variables. Top left: spatial, mtry=15; top right: spatial, mtry=22; bottom left: mtry=3; bottom right: mtry=5.

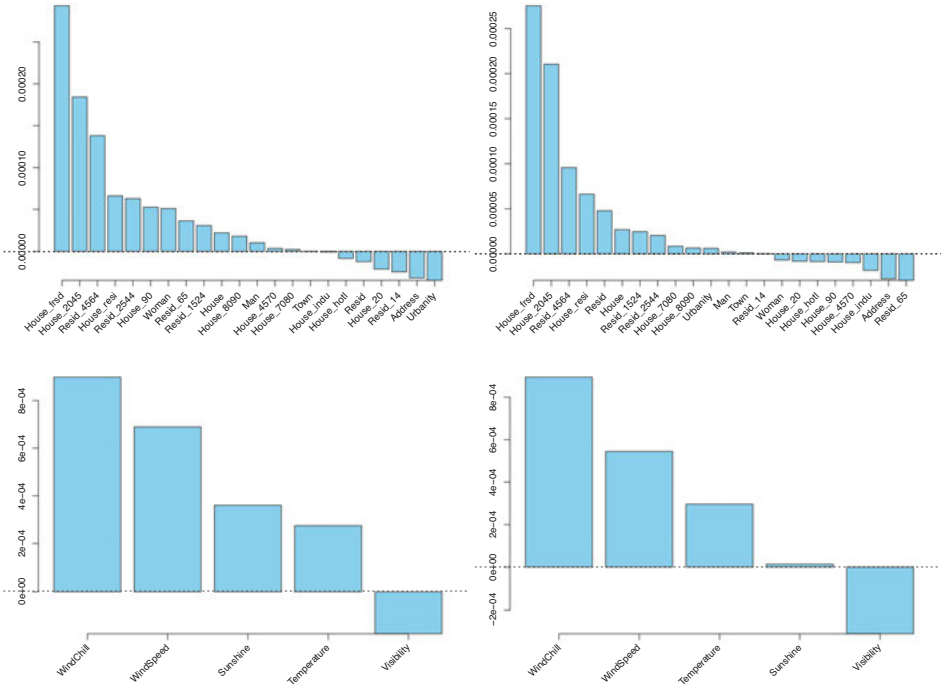


Figure 2.15 Tests on the influence of the hyperparameter – random seed – in the selection of explanatory variables. Top left: spatial, seed=30; top right: spatial, seed=200; bottom left: seed=30; bottom right: seed=200.

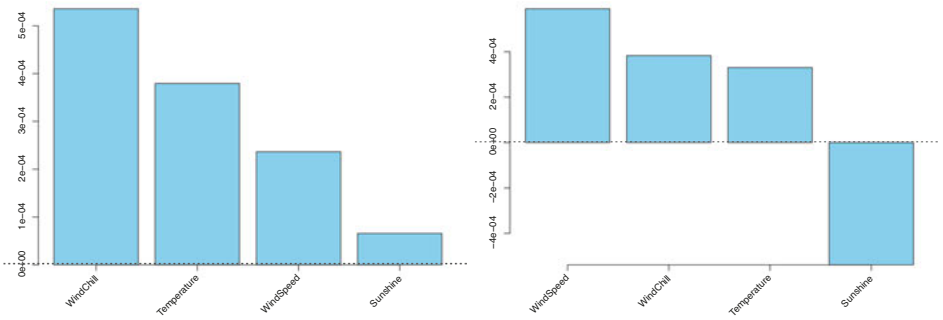


Figure 2.16 Tests on the influence of the weather variation in the selection of explanatory variables. Left: Heino; right:Hupsel. Note that the variable. visibility, is not accessible in these two weather stations.

CHAPTER 3

Infill Asymptotics for Logistic Regression Estimators for Spatio-temporal Point Processes

3.1 Introduction

In Chapter 2, we developed a heterogeneous spatio-temporal Poisson point process model to estimate the risk of chimney fire occurrences in the Twente region of the Netherlands. The intensity function we proposed is log-linear in a number of explanatory covariates. Similar model structures have been widely used, for instance, in Møller & Waagepetersen (2007) and Coeurjolly & Møller (2014).

To estimate the parameters of such a Poisson point process model, one may consider maximum likelihood estimation (e.g., Kutoyants, 1998; Møller & Waagepetersen, 2004). In practice, this approach involves an integral that needs to be approximated numerically, e.g. using quadrature points (Berman & Turner, 1992). For a non-Poisson point process model, the likelihood – if available at all – involves an intractable normalizing constant that must be approximated by cluster expansions (e.g., Ogata & Tanemura, 1981) or Markov chain Monte Carlo simulations (e.g., Møller & Waagepetersen, 2004), which are computationally expensive. Therefore, alternative approaches have been developed, such as minimum contrast estimation (e.g., Guyon, 1995) and, particularly, M-estimators based on unbiased estimating equations (e.g., Van der Vaart, 1998; Sørensen, 1999). Early examples of the latter approach are Takacs and Fiksel's pioneering work in the 1980s (Fiksel, 1984; Takacs, 1986; Fiksel, 1988) and maximum pseudo-likelihood estimation (Besag, 1978; Baddeley & Turner, 2000) for Gibbs and Markov point process models (e.g. Van Lieshout, 2000; Georgii, 2011). More recent methods include Poisson likelihood estimation (Schoenberg, 2005), logistic regression estimation (Baddeley et al., 2014) and the quasi-likelihood approach (Guan, Jalilian & Waagepetersen, 2015). Typically, these estimators are based on either the Campbell–Mecke theorem or the Nguyen–Zessin theorem (see, e.g., Daley & Vere-Jones, 2009). In Chapter 2, we opt to use logistic regression estimation because – in contrast to the quasi-likelihood approach - it is easy to implement using standard software for generalized linear models (Baddeley, Rubak & Turner, 2015), and because – other than Poisson likelihood estimation – it

involves a so-called dummy point process which can be tuned to focus on salient regions and times efficiently.

From a theoretical perspective, it is important to analyze the asymptotic properties of parameter estimators in order to establish statistical inference for point process models. In the current context, two asymptotic regimes can be formulated: increasing-domain asymptotics and infill asymptotics (Ripley, 2012). In the former, the observation window grows; in the latter, the window remains fixed but contains more and more points. A hybrid framework that combines the two regimes can be considered as well (e.g., Kutoyants, 1998).

For Poisson likelihood estimators, a vast body of asymptotic theory exists. Assuming that the point process of interest is Poisson, Rathbun & Cressie (1994) proved increasing-domain asymptotics and Kutoyants (1998) gave a detailed and comprehensive account in the hybrid framework. For general point processes, Schoenberg (2005) provided simple but sufficient conditions for consistency in the increasing-domain regime; Waagepetersen (2007) studied infill asymptotics for a class of clustered point processes models. For modified Poisson likelihood estimators designed to handle partially observed covariates, Rathbun, Shiffman & Gwaltney (2007) developed increasing-domain asymptotics for Poisson point processes in space and time; infill asymptotics for cluster point processes using similar estimators were treated in Waagepetersen (2008). Recently, for variable selection purposes, Thurman et al. (2015) considered increasing-domain asymptotics for regularized versions of Poisson likelihood estimators, whereas Choiruddin, Coeurjolly & Letué (2021) worked in the hybrid framework. Similar results for mis-specified models were also derived by Choiruddin, Coeurjolly & Waagepetersen (2020) for model selection purposes.

In contrast, the literature for logistic regression estimators is more limited. Rathbun (2013) discussed asymptotic theory for Poisson point processes in the increasing-domain regime. Baddeley et al. (2014) developed increasing-domain asymptotics for stationary spatial Gibbs point processes. Afterwards, Choiruddin, Coeurjolly & Letué (2018) proved asymptotic normality for general point process models but also in the increasing-domain framework. In this chapter, we study consistency and asymptotic normality for logistic regression estimators for spatio-temporal point processes in the infill regime and extend our central limit theorem to the estimators from general unbiased estimating equations that are based on the Campbell–Mecke theorem. Note that the infill framework is quite natural for Poisson point patterns observed in a fixed window; it is also suitable for general point patterns when replications are available (e.g., Pawlas, 2011; Chiu et al., 2013; Van Lieshout, 2021). In practice, replications can be obtained by observing point patterns within fixed domains over different time windows or by dividing a large-scale point pattern into independent, identically distributed pieces at various locations.

Specifically, our contributions are three-fold. First, we develop infill asymptotic theory for logistic regression estimators. Our asymptotic regime is based on independent and identically distributed point processes (cf., Section 2.4), dif-

fering from Waagepetersen (2008), who grew the intensity function by a factor of n. Moreover, other than in most of the literature, we give direct, explicit and self-contained proofs and provide sufficient conditions that are easy to verify in practice. Second, our asymptotic results enable us to improve the efficiency of logistic regression estimation by using regular dummy point processes. We propose consistent estimators for the asymptotic covariance matrix and demonstrate its practical use in a simulation study. Third, we extend our central limit theorem to the estimators obtained from general unbiased estimating equations, paving the way for further studies on the design of test functions.

The remainder of this chapter is organized as follows. Section 3.2 introduces the background and notation. Sections 3.3 and 3.4 present infill asymptotic results for logistic regression estimators for Poisson and general point process models, respectively. Section 3.5 proposes consistent estimators for the asymptotic covariance matrix and provides a simulation study. Section 3.6 discusses the extension to general unbiased estimating equations. Finally, the chapter finishes with a summary.

3.2 Background and Notation

3.2.1 Parametric Intensity Function

Let X be a spatio-temporal point process defined on a bounded non-empty open set $\mathcal{W} \times \mathcal{T} \subset \mathbb{R}^2 \times \mathbb{R}$ with intensity function $\lambda(\boldsymbol{w}, t)$. For ease of notation, we will write $\boldsymbol{s} = (\boldsymbol{w}, t) \in \mathcal{W} \times \mathcal{T}$. In this chapter, we assume that λ is of log-linear form offset by a measurable function b and parameterized by a vector $\boldsymbol{\theta}$ in some parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^p$:

$$\lambda(\boldsymbol{s}; \boldsymbol{\theta}) = b(\boldsymbol{s}) \exp\left[\boldsymbol{\theta}^\top \boldsymbol{z}(\boldsymbol{s})\right], \tag{3.1}$$

where $b(\boldsymbol{s}) : \mathcal{W} \times \mathcal{T} \to [0, \infty)$ is a measurable function that serves as the baseline or reference intensity, $\boldsymbol{z}(\boldsymbol{s}) = [z_1(\boldsymbol{s}), \ldots, z_p(\boldsymbol{s})]^\top : \mathcal{W} \times \mathcal{T} \to \mathbb{R}^p$ is an p-dimensional measurable vector of spatial and temporal covariates, and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^\top$ is the parameter vector. The gradient vector of $\lambda(\boldsymbol{s}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ then takes the form

$$\nabla\lambda(\boldsymbol{s}; \boldsymbol{\theta}) = \lambda(\boldsymbol{s}; \boldsymbol{\theta})\boldsymbol{z}(\boldsymbol{s}). \tag{3.2}$$

Conditions must be imposed on b and $\boldsymbol{z}$ to ensure that λ is integrable. In the sequel, it will sometimes also be necessary to assume that b, and thus λ, is strictly positive. When this is the case, we will state it explicitly.

3.2.2 Logistic Regression Estimation

Unbiased estimating equations for the parameters of a spatio-temporal point process model in general and logistic regression estimation in particular are based on the Campbell–Mecke theorem (see, e.g., Daley & Vere-Jones, 2009).

The theorem states that, for any real-valued measurable function f defined on $\mathcal{W} \times \mathcal{T}$ such that $f\lambda$ is absolutely integrable,

$$\mathbb{E}\left[\sum_{\boldsymbol{x} \in X} f(\boldsymbol{x})\right] = \int_{\mathcal{W}\times\mathcal{T}} f(\boldsymbol{s})\lambda(\boldsymbol{s})\mathrm{d}\boldsymbol{s}, \tag{3.3}$$

where $\boldsymbol{x}$ runs through the points of X. When λ is parameterized by a vector $\boldsymbol{\theta}$ as $\lambda(\boldsymbol{s};\boldsymbol{\theta})$, (3.3) provides a basis for estimating $\boldsymbol{\theta}$.

The logistic regression estimation is based on the vector function

$$\boldsymbol{f}(\boldsymbol{s};\boldsymbol{\theta}) = \nabla \log\left[\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta})}{\lambda(\boldsymbol{s};\boldsymbol{\theta}) + \rho(\boldsymbol{s})}\right] = \frac{\rho(\boldsymbol{s})/\lambda(\boldsymbol{s};\boldsymbol{\theta})}{\lambda(\boldsymbol{s};\boldsymbol{\theta}) + \rho(\boldsymbol{s})}\nabla\lambda(\boldsymbol{s};\boldsymbol{\theta}). \tag{3.4}$$

Here, one assumes that $\lambda(\boldsymbol{s};\boldsymbol{\theta})$ is a positive-valued differentiable function such that its gradient vector $\nabla\lambda(\boldsymbol{s};\boldsymbol{\theta})$ is absolutely integrable, and that $\rho(\boldsymbol{s})$ is a positive-valued measurable function also defined on $\mathcal{W} \times \mathcal{T}$. The idea is then to estimate both sides of (3.3) and solve the equations for $\boldsymbol{\theta}$. In order to approximate the right-hand side, one may use a 'dummy' point process D on $\mathcal{W} \times \mathcal{T}$ which is independent of X and has integrable intensity function ρ. Applying the Campbell–Mecke theorem to D, one finds that

$$\sum_{\boldsymbol{x} \in D} \frac{1}{\lambda(\boldsymbol{x};\boldsymbol{\theta}) + \rho(\boldsymbol{x})}\nabla\lambda(\boldsymbol{x};\boldsymbol{\theta}) \tag{3.5}$$

is an unbiased estimator for the right-hand side of (3.3) with $\boldsymbol{f}$ as in (3.4). Hence,

$$\begin{aligned}\boldsymbol{s}(X, D;\boldsymbol{\theta}) = &\sum_{\boldsymbol{x} \in X} \frac{\rho(\boldsymbol{x})/\lambda(\boldsymbol{x};\boldsymbol{\theta})}{\lambda(\boldsymbol{x};\boldsymbol{\theta}) + \rho(\boldsymbol{x})}\nabla\lambda(\boldsymbol{x};\boldsymbol{\theta}) \\ &- \sum_{\boldsymbol{x} \in D} \frac{1}{\lambda(\boldsymbol{x};\boldsymbol{\theta}) + \rho(\boldsymbol{x})}\nabla\lambda(\boldsymbol{x};\boldsymbol{\theta}) = \mathbf{0}\end{aligned} \tag{3.6}$$

is an unbiased estimating equation. It is interesting to observe that the middle part of (3.6) is exactly the gradient of a logistic log-likelihood function

$$L(X, D;\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in X} \log\left[\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta})}{\lambda(\boldsymbol{x};\boldsymbol{\theta}) + \rho(\boldsymbol{x})}\right] + \sum_{\boldsymbol{x} \in D} \log\left[\frac{\rho(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta}) + \rho(\boldsymbol{x})}\right]. \tag{3.7}$$

Thus, (3.6) can also be interpreted as a score function and solved using standard software for logistic regression for exponential family models such as the R-package *stats* (Venables & Ripley, 2002). The existence and uniqueness of a maximizer of (3.7) are ensured under proper conditions (Silvapulle, 1981).

For convenience, in the remainder of this chapter, we write $\boldsymbol{s}(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$ for $\boldsymbol{s}(X, D;\boldsymbol{\theta})$ and $L(X, D;\boldsymbol{\theta})$ to suppress the dependence on the point patterns X and D. Moreover, we use $\boldsymbol{\theta_0}$ and $\widehat{\boldsymbol{\theta}}$ to, respectively, denote the true value and the estimator of $\boldsymbol{\theta}$.

3.2.3 First Two Moments of First-order U-statistics

Following Reitzner & Schulte (2013), we call random vectors of the form

$$\boldsymbol{H} = \sum_{\boldsymbol{x} \in X} [h_1(\boldsymbol{x}), \ldots, h_p(\boldsymbol{x})]^\top \tag{3.8}$$

first-order U-statistics of X if every $h_l\lambda$ is absolutely integrable.

Suppose that X also has second-order product density function $\lambda^{(2)}$. By the Campbell–Mecke theorem and in analogy to (3.3),

$$\mathbb{E}\left[\sum_{\boldsymbol{x} \in X} \sum_{\boldsymbol{y} \in X, \boldsymbol{y} \neq \boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\right] = \int \int_{(\mathcal{W} \times \mathcal{T})^2} f(\boldsymbol{s}, \boldsymbol{u}) \lambda^{(2)}(\boldsymbol{s}, \boldsymbol{u}) \mathrm{d}\boldsymbol{s} \mathrm{d}\boldsymbol{u} \tag{3.9}$$

for any real-valued measurable function f defined on $(\mathcal{W} \times \mathcal{T})^2$ such that $f\lambda^{(2)}$ is absolutely integrable. Then, they can be used to derive the first two moments of $\boldsymbol{H}$. Indeed,

$$\mathbb{E}(\boldsymbol{H}) = \left[\int_{\mathcal{W} \times \mathcal{T}} h_l(\boldsymbol{s}) \lambda(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}\right]_{l=1}^{p} \tag{3.10}$$

and

$$\begin{aligned} \mathbb{E}\left(\boldsymbol{H}\boldsymbol{H}^\top\right) = &\left[\int_{\mathcal{W} \times \mathcal{T}} h_k(\boldsymbol{s}) h_l(\boldsymbol{s}) \lambda(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p} \\ &+ \left[\int \int_{(\mathcal{W} \times \mathcal{T})^2} h_k(\boldsymbol{s}) h_l(\boldsymbol{u}) \lambda^{(2)}(\boldsymbol{s}, \boldsymbol{u}) \mathrm{d}\boldsymbol{s} \mathrm{d}\boldsymbol{u}\right]_{k,l=1}^{p}. \end{aligned} \tag{3.11}$$

If $\mathbb{E}[\boldsymbol{H}\boldsymbol{H}^\top]$ is finite and $\lambda > 0$, the entries in the covariance matrix of $\boldsymbol{H}$ are finite and can be expressed in terms of the pair correlation function $g(\boldsymbol{s}, \boldsymbol{u})$ of X as

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{H}) = &\left[\int_{\mathcal{W} \times \mathcal{T}} h_k(\boldsymbol{s}) h_l(\boldsymbol{s}) \lambda(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p} \\ &+ \left[\int \int_{(\mathcal{W} \times \mathcal{T})^2} h_k(\boldsymbol{s}) h_l(\boldsymbol{u}) [g(\boldsymbol{s}, \boldsymbol{u}) - 1] \lambda(\boldsymbol{s}) \lambda(\boldsymbol{u}) \mathrm{d}\boldsymbol{s} \mathrm{d}\boldsymbol{u}\right]_{k,l=1}^{p}. \end{aligned} \tag{3.12}$$

3.2.4 Infill Asymptotic Regime

The infill asymptotic regime considered in this study is as defined in Pawlas (2011), Chiu et al. (2013) and Van Lieshout (2021).

Let $\{Y_i\}$ and $\{E_i\}$, with $i \in \mathbb{N}^+$, be two independent sequences of independent and identically distributed spatio-temporal point processes with intensity functions λ and ρ, respectively. Set

$$X_n = \bigcup_{i=1}^{n} Y_i, \quad D_n = \bigcup_{i=1}^{n} E_i. \tag{3.13}$$

Write λ_n for the intensity function of X_n and ρ_n for that of the 'dummy' point process D_n. Thus $\lambda_n = n\lambda$ and $\rho_n = n\rho$. We assume that the intensity functions of X_n and D_n increase at the same rate. If λ is of the form (3.1), for any $n \in \mathbb{N}^+$, the score function (3.6) based on X_n and D_n becomes

$$\boldsymbol{s}_n(\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in X_n} \frac{\rho(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta}) + \rho(\boldsymbol{x})} \boldsymbol{z}(\boldsymbol{x}) - \sum_{\boldsymbol{x} \in D_n} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta})}{\lambda(\boldsymbol{x};\boldsymbol{\theta}) + \rho(\boldsymbol{x})} \boldsymbol{z}(\boldsymbol{x}). \tag{3.14}$$

Note that the terms in the two sums above do not depend on n while the subscript $\boldsymbol{x}$ runs through the points of X_n and D_n. Taking the limit as $n \to \infty$, one obtains an asymptotic regime that Ripley (2012) calls 'infill asymptotics'.

Intuitively, under this regime, the estimate for the parameter vector $\boldsymbol{\theta}$ will become more precise when the points observed in the fixed window $\mathcal{W} \times \mathcal{T}$ become more dense. Our aim in the remainder of this chapter is to analyze the asymptotic behaviour of the estimator $\widehat{\boldsymbol{\theta}}_n$ defined by (3.14) as $n \to \infty$.

3.3 Infill Asymptotics for Poisson Point Process Models

In this section, we derive infill asymptotics for logistic regression estimators in the case that the point process of interest is Poisson. As discussed in Section 2.4.5, it is motivated from the fire prediction study, where the Poisson point process of interest existed in a fixed window only.

It is worth noting that the probability distribution of a Poisson point process is completely specified by its first-order moment measure, cf. Section 1.1. The assumption of a log-linear intensity function parameterized as (3.1) is natural and constitutes an exponential family with the components of $\sum_{\boldsymbol{x} \in X} \boldsymbol{z}(\boldsymbol{x})$ as sufficient statistics. Moreover, the pair correlation function of a Poisson process is always equal to one. Thus, the infill asymptotic regime defined in Section 3.2.4 is particularly appropriate due to the conditional independence of points given a fixed number.

For ease of reference, we list the following conditions that are required to derive the asymptotic results.

(C1) $\{Y_i\}$ and $\{E_i\}$ with $i \in \mathbb{N}^+$ are two independent sequences of independent and identically distributed spatio-temporal point processes on some bounded open set $\mathcal{W} \times \mathcal{T} \subset \mathbb{R}^2 \times \mathbb{R}$ and defined on the same underlying probability space. Set $X_n = \cup_{i=1}^n Y_i$ and $D_n = \cup_{i=1}^n E_i$.

(C2) Y_i is a Poisson point process with intensity function $\lambda(\boldsymbol{s};\boldsymbol{\theta})$ given by (3.1), where $b > 0$ is an integrable function, $\boldsymbol{z}$ is a measurable vector of covariates and the parameter vector $\boldsymbol{\theta}$ lies in an open set $\boldsymbol{\Theta} \subset \mathbb{R}^p$; E_i has integrable intensity function $\rho(\boldsymbol{s}) > 0$.

(C3) E_i has bounded pair correlation function $g(\boldsymbol{s}, \boldsymbol{u})$: $\sup_{(\boldsymbol{s},\boldsymbol{u}) \in (\mathcal{W}\times\mathcal{T})^2} g(\boldsymbol{s}, \boldsymbol{u}) < \infty$.

(C4) For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there exist $\epsilon_1(\boldsymbol{\theta}), \epsilon_2(\boldsymbol{\theta}) > 0$ such that $\epsilon_1(\boldsymbol{\theta}) < \inf_{\boldsymbol{s} \in \mathcal{W} \times \mathcal{T}} \rho(\boldsymbol{s}) / \lambda(\boldsymbol{s}; \boldsymbol{\theta})$ and $\sup_{\boldsymbol{s} \in \mathcal{W} \times \mathcal{T}} \rho(\boldsymbol{s}) / \lambda(\boldsymbol{s}; \boldsymbol{\theta}) < \epsilon_2(\boldsymbol{\theta})$.

(C5) The elements of the measurable covariate vector $\boldsymbol{z}$ are all bounded: $\sup_{\boldsymbol{s} \in \mathcal{W} \times \mathcal{T}} \|\boldsymbol{z}(\boldsymbol{s})\| < \infty$.

(C6) The parameter space $\boldsymbol{\Theta}$ is convex.

(C7) The parametric model for λ is identifiable: $\lambda(\boldsymbol{s}; \boldsymbol{\theta}) = \lambda(\boldsymbol{s}; \widetilde{\boldsymbol{\theta}})$ almost everywhere on $\mathcal{W} \times \mathcal{T}$ implies $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$.

(C8) The $[p \times p]$-dimensional matrix $\boldsymbol{s}$, whose (k, l)-th entry reads $\int_{\mathcal{W} \times \mathcal{T}} \lambda(\boldsymbol{s}; \boldsymbol{\theta_0}) \rho(\boldsymbol{s})\ z_k(\boldsymbol{s}) z_l(\boldsymbol{s}) / [\lambda(\boldsymbol{s}; \boldsymbol{\theta_0}) + \rho(\boldsymbol{s})] \mathrm{d}\boldsymbol{s}$ with $\boldsymbol{\theta_0} \in \boldsymbol{\Theta}$, is positive definite.

A few remarks on some of the conditions are appropriate. In condition (C2), Y_i is assumed to be a Poisson point process, however, this assumption will be relaxed in Section 3.4. In condition (C3), E_i is not required to be a Poisson point process, as its realizations are only used to approximate an integral. Condition (C4) is reasonable, recalling the rule of thumb recommended by Baddeley et al. (2014) for selecting ρ such that $\rho \simeq 4\lambda$. Condition (C7) is necessary for strong consistency (cf., Theorem 3.2). Condition (C8) helps ensure the attainment of the estimator $\widehat{\boldsymbol{\theta}}_n$ as $n \to \infty$ (cf., Theorem 3.9).

In the remainder, we will use $P_{\boldsymbol{\theta_0}}$ to denote the distribution of (X_n, D_n) under the true parameter value $\boldsymbol{\theta_0}$.

3.3.1 Strong Consistency

To establish strong consistency, we start our investigations with the asymptotic behaviour of the scaled log-likelihood function.

Lemma 3.1. *Assume that conditions (C1)–(C2) and (C4)–(C5) hold. Define $L_n(\boldsymbol{\theta}) = L(X_n, D_n; \boldsymbol{\theta})$ by (3.7) with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then, as $n \to \infty$, $L_n(\boldsymbol{\theta})/n$ converges $P_{\boldsymbol{\theta_0}}$-almost surely to*

$$\int_{\mathcal{W} \times \mathcal{T}} \left\{ \lambda(\boldsymbol{s}; \boldsymbol{\theta_0}) \log \left[\frac{\lambda(\boldsymbol{s}; \boldsymbol{\theta})}{\lambda(\boldsymbol{s}; \boldsymbol{\theta}) + \rho(\boldsymbol{s})} \right] + \rho(\boldsymbol{s}) \log \left[\frac{\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s}; \boldsymbol{\theta}) + \rho(\boldsymbol{s})} \right] \right\} \mathrm{d}\boldsymbol{s}.$$

Proof. Under conditions (C1)–(C2) and recalling the logistic log-likelihood function (3.7),

$$\frac{L_n(\boldsymbol{\theta})}{n} = \frac{1}{n} \sum_{\boldsymbol{x} \in X_n} \log \left[\frac{\lambda(\boldsymbol{x}; \boldsymbol{\theta})}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})} \right] + \frac{1}{n} \sum_{\boldsymbol{x} \in D_n} \log \left[\frac{\rho(\boldsymbol{x})}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})} \right],$$

which consists of two first-order U-statistics defined on X_n and D_n. We derive their strong convergence separately.

Write the first-order U-statistic defined on X_n as

$$\frac{1}{n} \sum_{\boldsymbol{x} \in X_n} \log \left[\frac{\lambda(\boldsymbol{x}; \boldsymbol{\theta})}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})} \right] = \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{\boldsymbol{x} \in Y_i} \log \left[\frac{\lambda(\boldsymbol{x}; \boldsymbol{\theta})}{\lambda(\boldsymbol{x}; \boldsymbol{\theta}) + \rho(\boldsymbol{x})} \right] \right\}.$$

The sum is the average of independent and identically distributed real-valued random variables. By the Campbell–Mecke theorem,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta_0}}\left\{\sum_{\boldsymbol{x}\in Y_i}\log\left[\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta})}{\lambda(\boldsymbol{x};\boldsymbol{\theta})+\rho(\boldsymbol{x})}\right]\right\} = \\ \int_{\mathcal{W}\times\mathcal{T}}\log\left[\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})+\rho(\boldsymbol{s})}\right]\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\mathrm{d}\boldsymbol{s}.\end{aligned} \tag{3.15}$$

Conditions (C2) and (C5) imply that the intensity function $\lambda(\boldsymbol{s};\boldsymbol{\theta_0})$ is integrable on $\mathcal{W}\times\mathcal{T}$. By conditions (C2) and (C4),

$$\log\left[\frac{1}{1+\epsilon_2(\boldsymbol{\theta})}\right] < \log\left[\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})+\rho(\boldsymbol{s})}\right] < \log\left[\frac{1}{1+\epsilon_1(\boldsymbol{\theta})}\right].$$

Thus, the $P_{\boldsymbol{\theta_0}}$-mean in (3.15) is finite for all $\boldsymbol{\theta}$. Kolmogorov's strong law of large numbers implies that the first-order U-statistic defined on X_n converges $P_{\boldsymbol{\theta_0}}$-almost surely to the integral in the right-hand side of (3.15).

Similarly, for the first-order U-statistic defined on D_n, as $n\to\infty$,

$$\begin{aligned}\frac{1}{n}\sum_{\boldsymbol{x}\in D_n}\log\left[\frac{\rho(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta})+\rho(\boldsymbol{x})}\right] &\overset{P_{\boldsymbol{\theta_0}}-a.s.}{\to} \mathbb{E}_{\boldsymbol{\theta_0}}\left\{\sum_{\boldsymbol{x}\in E_i}\log\left[\frac{\rho(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta})+\rho(\boldsymbol{x})}\right]\right\} \\ &= \int_{\mathcal{W}\times\mathcal{T}}\log\left[\frac{\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})+\rho(\boldsymbol{s})}\right]\rho(\boldsymbol{s})\mathrm{d}\boldsymbol{s}\end{aligned}$$

because $\rho(\boldsymbol{s})$ is integrable by condition (C2) and

$$\log\left[\frac{\epsilon_1(\boldsymbol{\theta})}{1+\epsilon_1(\boldsymbol{\theta})}\right] < \log\left[\frac{\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})+\rho(\boldsymbol{s})}\right] < \log\left[\frac{\epsilon_2(\boldsymbol{\theta})}{1+\epsilon_2(\boldsymbol{\theta})}\right]$$

by conditions (C2) and (C4).

The proof is completed by combining the two strong convergence results. □

Recall that the logistic regression estimator minimizes $U_n(\boldsymbol{\theta}) = -L_n(\boldsymbol{\theta})/n$ and thus can be considered as a minimum contrast estimator. The next theorem is concerned with strong consistency.

Theorem 3.2. *Assume that conditions (C1)–(C2) and (C4)–(C7) hold. Define* $L_n(\boldsymbol{\theta}) = L(X_n, D_n;\boldsymbol{\theta})$ *by (3.7) with* $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ *and set* $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta})$. *If* $\widehat{\boldsymbol{\theta}}_n$ *is attained, as* $n\to\infty$, $\widehat{\boldsymbol{\theta}}_n$ *converges* $P_{\boldsymbol{\theta_0}}$*-almost surely to* $\boldsymbol{\theta_0}$.

Proof. Conditions (C2) and (C6) ensure that the parameter space $\boldsymbol{\Theta}$ is open and convex.

First, we prove that, for almost every sample in the underlying probability space, the realizations of the function $\boldsymbol{\theta}\mapsto U_n(\boldsymbol{\theta})$ are convex. By condition (C2), the intensity function $\lambda(\boldsymbol{x};\boldsymbol{\theta})$ has the log-linear form (3.1) and is thus twice

differentiable with respect to the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then, the Hessian matrix of $-L_n(\boldsymbol{\theta})$ reads

$$\left[\sum_{\boldsymbol{x} \in X_n \cup D_n} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta})\rho(\boldsymbol{x})}{\left[\lambda(\boldsymbol{x};\boldsymbol{\theta}) + \rho(\boldsymbol{x})\right]^2} z_k(\boldsymbol{x}) z_l(\boldsymbol{x}) \right]_{k,l=1}^{p}.$$

By decomposition, it can be written into the product of a matrix $\boldsymbol{M}$ and its transpose as $\boldsymbol{M}^\top \boldsymbol{M}$, where $\boldsymbol{M}$ is a $[|X_n \cup D_n| \times p]$-dimensional matrix given by

$$\boldsymbol{M} = \left[\frac{\left[\lambda(\boldsymbol{x}_i;\boldsymbol{\theta})\rho(\boldsymbol{x}_i)\right]^{1/2}}{\lambda(\boldsymbol{x}_i;\boldsymbol{\theta}) + \rho(\boldsymbol{x}_i)} z_k(\boldsymbol{x}_i) \right]_{i=1,k=1}^{|X_n \cup D_n|,p}.$$

Here, $|X_n \cup D_n|$ denotes the number of points in $X_n \cup D_n$, the subscript i runs through all points in $X_n \cup D_n$ and k runs through p covariates. One can readily obtain that the Hessian matrix of $-L_n(\boldsymbol{\theta})$ is positive semi-definite for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which implies that $-L_n(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, and thus $-L_n(\boldsymbol{\theta})/n$, is convex.

Secondly, we prove that, as $n \to \infty$, $U_n(\boldsymbol{\theta}) - U_n(\boldsymbol{\theta_0})$ converges $P_{\boldsymbol{\theta_0}}$-almost surely to a function $K(\boldsymbol{\theta}, \boldsymbol{\theta_0})$ which is non-negative and vanishes only at $\boldsymbol{\theta} = \boldsymbol{\theta_0}$. By Lemma 3.1, under conditions (C1)–(C2) and (C4)–(C5), $U_n(\boldsymbol{\theta}) - U_n(\boldsymbol{\theta_0})$ converges $P_{\boldsymbol{\theta_0}}$-almost surely to

$$\int_{\mathcal{W}\times\mathcal{T}} \left\{ \lambda(\boldsymbol{s};\boldsymbol{\theta_0}) \log\left[\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})}\right] - \left[\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})\right] \log\left[\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta}) + \rho(\boldsymbol{s})}\right] \right\} \mathrm{d}\boldsymbol{s}.$$

Denote this limit by $K(\boldsymbol{\theta}, \boldsymbol{\theta_0})$ and the integrand by $k(\boldsymbol{s};\boldsymbol{\theta},\boldsymbol{\theta_0})$. Clearly, $K(\boldsymbol{\theta_0},\boldsymbol{\theta_0})$ $= 0$. Furthermore,

$$k(\boldsymbol{s};\boldsymbol{\theta},\boldsymbol{\theta_0}) = \lambda(\boldsymbol{s};\boldsymbol{\theta}) \times \left\{ \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})} \log\left[\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})}\right] - \left[\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})} + \frac{\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})}\right] \log\left[\frac{\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})} + \frac{\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})}}{1 + \frac{\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta})}}\right] \right\}.$$

By condition (C2), $\lambda(\boldsymbol{s};\boldsymbol{\theta}) > 0$. Consider the function $a \mapsto a\log a - (a+b)\log[(a+b)/(1+b)]$ with $a, b > 0$. Its derivative with respect to a is $\log[a(1+b)/(a+b)]$. The function is strictly decreasing when $a \in (0,1)$ and strictly increasing when $a \in (1, +\infty)$. For $a = 1$, $a\log a - (a+b)\log[(a+b)/(1+b)] = 0$. Thus, $k(\boldsymbol{s};\boldsymbol{\theta},\boldsymbol{\theta_0})$ is non-negative and is strictly positive when $\lambda(\boldsymbol{s};\boldsymbol{\theta}) \neq \lambda(\boldsymbol{s};\boldsymbol{\theta_0})$.

Under condition (C7), strong consistency then follows from an appeal to the Proposition below Guyon (1995, Theorem 3.4.4). □

Remark 3.3. *Note that the proofs of Lemma 3.1 and Theorem 3.2 do not depend on the assumption that Y_i, and thus X_n, is a Poisson point process. However, strong consistency relies on the concavity of the log-likelihood function of logistic regression estimation.*

3.3.2 Asymptotic Normality

To establish asymptotic normality, we start our investigations with the Taylor series of the score function (3.14).

For every component of the score function, denoted by $\boldsymbol{s}_{n,i}(\boldsymbol{\theta})$ with $1 \leq i \leq p$, the second-order Taylor expansion of $\boldsymbol{s}_{n,i}(\widehat{\boldsymbol{\theta}}_n)$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta_0}$ reads

$$\begin{aligned}\boldsymbol{s}_{n,i}(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{s}_{n,i}(\boldsymbol{\theta_0}) + \nabla \boldsymbol{s}_{n,i}(\boldsymbol{\theta_0})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}) \\ + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})^\top \nabla^2 \boldsymbol{s}_{n,i}(\boldsymbol{\theta'^{,i}})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}),\end{aligned} \tag{3.16}$$

where $\nabla \boldsymbol{s}_{n,i}(\boldsymbol{\theta})$ is the $[1 \times p]$-dimensional vector containing the first-order partial derivatives of $\boldsymbol{s}_{n,i}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and $\nabla^2 \boldsymbol{s}_{n,i}(\boldsymbol{\theta})$ is the $[p \times p]$-dimensional matrix containing the second-order partial derivatives of $\boldsymbol{s}_{n,i}(\boldsymbol{\theta})$. Moreover, $\boldsymbol{\theta'^{,i}}$ is a convex combination of $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta_0}$ which, by condition (C6), lies in $\boldsymbol{\Theta}$ as well.

Write $\nabla \boldsymbol{s}_n(\boldsymbol{\theta_0})$ for the matrix whose i-th row is $\nabla \boldsymbol{s}_{n,i}(\boldsymbol{\theta_0})$ and assume its inverse is well-defined. Heuristically, the idea is to equate (3.16) to zero, ignore the error term which is the quadratic form in (3.16), and rearrange the remaining terms to obtain

$$n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}) \approx \left[-\frac{\nabla \boldsymbol{s}_n(\boldsymbol{\theta_0})}{n}\right]^{-1} \frac{\boldsymbol{s}_n(\boldsymbol{\theta_0})}{n^{1/2}}. \tag{3.17}$$

In the next two lemmas, we study the asymptotic behaviours of the terms $-\nabla \boldsymbol{s}_n(\boldsymbol{\theta_0})/n$ and $\boldsymbol{s}_n(\boldsymbol{\theta_0})/n^{1/2}$, respectively.

Lemma 3.4. *Assume that conditions (C1)–(C2) and (C5) hold. Define* $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ *by (3.14) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. *Then, as* $n \to \infty$, $-\nabla \boldsymbol{s}_n(\boldsymbol{\theta_0})/n$ *converges* $P_{\boldsymbol{\theta_0}}$*-almost surely to*

$$\boldsymbol{U} = \left[\int_{\mathcal{W}\times\mathcal{T}} \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})} z_k(\boldsymbol{s}) z_l(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p}.$$

Proof. Under conditions (C1)–(C2) and recalling the score function (3.14),

$$-\frac{\nabla \boldsymbol{s}_n(\boldsymbol{\theta_0})}{n} = \left[\sum_{\boldsymbol{x}\in X_n \cup D_n} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})\rho(\boldsymbol{x})}{n\left[\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})\right]^2} z_k(\boldsymbol{x}) z_l(\boldsymbol{x})\right]_{k,l=1}^{p}.$$

To prove component-wise strong convergence, consider the (k,l)-th entry of the matrix above which, recalling condition (C1), is given by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{\boldsymbol{x}\in Y_i \cup E_i} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})\rho(\boldsymbol{x})}{[\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})]^2} z_k(\boldsymbol{x}) z_l(\boldsymbol{x})\right\}.$$

The sum is the average of independent and identically distributed real-valued random variables. By the Campbell–Mecke theorem,

$$\mathbb{E}_{\boldsymbol{\theta_0}}\left\{\sum_{\boldsymbol{x}\in Y_i \cup E_i} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})\rho(\boldsymbol{x}) z_k(\boldsymbol{x}) z_l(\boldsymbol{x})}{[\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})]^2}\right\} = \int_{\mathcal{W}\times\mathcal{T}} \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\rho(\boldsymbol{s}) z_k(\boldsymbol{s}) z_l(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})} \mathrm{d}\boldsymbol{s}.$$

By condition (C5), the covariate terms in the integrand are bounded. By condition (C2), $0 < \lambda(\boldsymbol{s};\boldsymbol{\theta_0})\rho(\boldsymbol{s})/[\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})] < \rho(\boldsymbol{s})$ and $\rho(\boldsymbol{s})$ is integrable. Kolmogorov's strong law of large numbers implies the claimed $P_{\boldsymbol{\theta_0}}$-almost sure convergence. □

Lemma 3.5. *Assume that conditions (C1)–(C3) and (C5) hold. Define* $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ *by (3.14) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. *Then, as* $n \to \infty$, $\boldsymbol{s}_n(\boldsymbol{\theta_0})/n^{1/2}$ *converges under* $P_{\boldsymbol{\theta_0}}$ *in distribution to a p-dimensional normally distributed random vector with mean zero and covariance matrix*

$$\boldsymbol{V} = \left[\int_{\mathcal{W}\times\mathcal{T}} \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})} z_k(\boldsymbol{s}) z_l(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p} + \left[\int\int_{(\mathcal{W}\times\mathcal{T})^2} \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})\rho(\boldsymbol{s})\rho(\boldsymbol{u}) z_k(\boldsymbol{s}) z_l(\boldsymbol{u})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})][\lambda(\boldsymbol{u};\boldsymbol{\theta_0}) + \rho(\boldsymbol{u})]} [g(\boldsymbol{s},\boldsymbol{u}) - 1] \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{u}\right]_{k,l=1}^{p}.$$

Proof. Under conditions (C1)–(C2) and recalling the score function (3.14),

$$\frac{\boldsymbol{s}_n(\boldsymbol{\theta_0})}{n^{1/2}} = \sum_{\boldsymbol{x}\in X_n} \frac{\rho(\boldsymbol{x})\boldsymbol{z}(\boldsymbol{x})}{n^{1/2}[\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})]} - \sum_{\boldsymbol{x}\in D_n} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})\boldsymbol{z}(\boldsymbol{x})}{n^{1/2}[\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})]}.$$

It consists of two first-order U-statistics defined on X_n and D_n. Write it as

$$\begin{aligned}\frac{\boldsymbol{s}_n(\boldsymbol{\theta_0})}{n^{1/2}} &= \frac{\boldsymbol{s}_n(X_n;\boldsymbol{\theta_0})}{n^{1/2}} - \frac{\boldsymbol{s}_n(D_n;\boldsymbol{\theta_0})}{n^{1/2}} \\ &= n^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\boldsymbol{x}\in Y_i} \frac{\rho(\boldsymbol{x})\boldsymbol{z}(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})}\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\boldsymbol{x}\in E_i} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})\boldsymbol{z}(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})}\right]\right\}.\end{aligned}$$

We discuss the two first-order U-statistics separately.

Consider $n^{-1/2}\boldsymbol{s}_n(X_n;\boldsymbol{\theta_0})$. Note that the term in the first curly bracket is the average of independent and identically distributed real-valued random vectors. By the Campbell–Mecke theorem and recalling Section 3.2.3,

$$\mathrm{M}_{\boldsymbol{\theta_0}} = \mathbb{E}_{\boldsymbol{\theta_0}}\left[\sum_{\boldsymbol{x}\in Y_i} \frac{\rho(\boldsymbol{x})\boldsymbol{z}(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})}\right] = \left[\int_{\mathcal{W}\times\mathcal{T}} \frac{\rho(\boldsymbol{s})\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) z_l(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})} \mathrm{d}\boldsymbol{s}\right]_{l=1}^{p}$$

and, because Y_i is a Poisson point process under condition (C2),

$$\mathrm{Cov}_{\boldsymbol{\theta_0}}\left[\sum_{\boldsymbol{x}\in Y_i} \frac{\rho(\boldsymbol{x})\boldsymbol{z}(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + \rho(\boldsymbol{x})}\right] = \left[\int_{\mathcal{W}\times\mathcal{T}} \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\rho^2(\boldsymbol{s}) z_k(\boldsymbol{s}) z_l(\boldsymbol{s})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})]^2} \mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p}. \tag{3.18}$$

By condition (C5), the covariate terms in the integrand are bounded. By condition (C2), $0 < \lambda(\boldsymbol{s};\boldsymbol{\theta_0})\rho^2(\boldsymbol{s})/[\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + \rho(\boldsymbol{s})]^2 < \lambda(\boldsymbol{s};\boldsymbol{\theta_0})$ and $\lambda(\boldsymbol{s};\boldsymbol{\theta_0})$ is integrable. The multi-variate Lindeberg–Lévy central limit theorem implies that $n^{-1/2}[\boldsymbol{s}_n(X_n;\boldsymbol{\theta_0}) - n\mathrm{M}_{\boldsymbol{\theta_0}}]$ converges under $P_{\boldsymbol{\theta_0}}$ in distribution to a p-dimensional

normally distributed random vector with mean zero and covariance matrix given by (3.18).

Analogously, by condition (C3), $n^{-1/2}[\boldsymbol{s}_n(D_n, \boldsymbol{\theta_0}) - n\mathrm{M}_{\boldsymbol{\theta_0}}]$ converges under $P_{\boldsymbol{\theta_0}}$ in distribution to a p-dimensional normally distributed random vector with mean zero and covariance matrix

$$\mathrm{Cov}_{\boldsymbol{\theta_0}}\left[\sum_{\boldsymbol{x}\in E_i}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})\boldsymbol{z}(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})+\rho(\boldsymbol{x})}\right] = \left[\int_{\mathcal{W}\times\mathcal{T}}\frac{\lambda^2(\boldsymbol{s};\boldsymbol{\theta_0})\rho(\boldsymbol{s})z_k(\boldsymbol{s})z_l(\boldsymbol{s})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})]^2}\mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p}$$
$$+\left[\int\int_{(\mathcal{W}\times\mathcal{T})^2}\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})\rho(\boldsymbol{s})\rho(\boldsymbol{u})z_k(\boldsymbol{s})z_l(\boldsymbol{u})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})][\lambda(\boldsymbol{u};\boldsymbol{\theta_0})+\rho(\boldsymbol{u})]}[g(\boldsymbol{s},\boldsymbol{u})-1]\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{u}\right]_{k,l=1}^{p}.$$

By condition (C2),

$$0 < \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})\rho(\boldsymbol{s})\rho(\boldsymbol{u})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})][\lambda(\boldsymbol{u};\boldsymbol{\theta_0})+\rho(\boldsymbol{u})]} < \lambda(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})$$

and $\lambda(\boldsymbol{s};\boldsymbol{\theta_0}), \lambda(\boldsymbol{u};\boldsymbol{\theta_0})$ are integrable.

Applying Lévy's continuity theorem and using the independence of X_n and D_n under condition (C1), the weak limits for the two first-order U-statistics can be combined, which completes the proof. □

Remark 3.6. *The pair correlation function $g(\boldsymbol{s},\boldsymbol{u})$ of E_i can be tuned to control the covariance matrix $\boldsymbol{V}$. When E_i, and thus D_n, is a Poisson point process, $g(\boldsymbol{s},\boldsymbol{u}) \equiv 1$ and the second term in $\boldsymbol{V}$ vanishes. Moreover, considering regular point processes E_i, whose $g(\boldsymbol{s},\boldsymbol{u}) < 1$, may be useful to reduce the variance.*

Recalling (3.17), we are now in a position to conjecture a central limit theorem for logistic regression estimators. For a formal proof, though, we need to analyze the error term in the Taylor series (3.16). The next theorem is concerned with asymptotic normality.

Theorem 3.7. *Assume that conditions (C1)–(C8) hold. Define $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ by (3.14) with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and let $\widehat{\boldsymbol{\theta}}_n$ be the estimator for which $\boldsymbol{s}_n(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$. If $\widehat{\boldsymbol{\theta}}_n$ is attained, as $n \to \infty$, $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})$ converges under $P_{\boldsymbol{\theta_0}}$ in distribution to a p-dimensional normally distributed random vector with mean zero and covariance matrix $\boldsymbol{U}^{-1}\boldsymbol{V}(\boldsymbol{U}^{-1})^\top$, where $\boldsymbol{U}, \boldsymbol{V}$ are as defined in Lemma 3.4 and Lemma 3.5.*

Proof. Consider the error term in the Taylor series (3.16) which, under conditions (C1) and (C2) and recalling the score function (3.14), reads

$$\frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})^\top \nabla^2 \boldsymbol{s}_{n,i}(\boldsymbol{\theta'^{,i}})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}) = \frac{1}{2}\sum_{k=1}^{p}\sum_{l=1}^{p}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})_k(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})_l \times$$
$$\left\{\sum_{\boldsymbol{x}\in X_n\cup D_n}\frac{\left[\lambda(\boldsymbol{x};\boldsymbol{\theta'^{,i}}) - \rho(\boldsymbol{x})\right]\lambda(\boldsymbol{x};\boldsymbol{\theta'^{,i}})\rho(\boldsymbol{x})z_i(\boldsymbol{x})z_k(\boldsymbol{x})z_l(\boldsymbol{x})}{\left[\lambda(\boldsymbol{x};\boldsymbol{\theta'^{,i}}) + \rho(\boldsymbol{x})\right]^3}\right\}$$

for all choices of $\boldsymbol{\theta'^{,i}}$ among the convex combinations of $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta_0}$.

First of all, note that

$$\sum_{\boldsymbol{x}\in X_n\cup D_n} \frac{\left[\lambda(\boldsymbol{x};\boldsymbol{\theta'^{,i}})-\rho(\boldsymbol{x})\right]\lambda(\boldsymbol{x};\boldsymbol{\theta'^{,i}})\rho(\boldsymbol{x})z_i(\boldsymbol{x})z_k(\boldsymbol{x})z_l(\boldsymbol{x})}{n\left[\lambda(\boldsymbol{x};\boldsymbol{\theta'^{,i}})+\rho(\boldsymbol{x})\right]^3} \tag{3.19}$$

is bounded in absolute value by

$$G_{n,ikl}=\sum_{\boldsymbol{x}\in X_n\cup D_n}\frac{|z_i(\boldsymbol{x})z_k(\boldsymbol{x})z_l(\boldsymbol{x})|}{n}.$$

This bound does not depend on $\boldsymbol{\theta'^{,i}}$ and thus does not depend on $\widehat{\boldsymbol{\theta}}_n$. By the Campbell–Mecke theorem,

$$\mathbb{E}_{\boldsymbol{\theta_0}}\left(G_{n,ikl}\right)=\int_{\mathcal{W}\times\mathcal{T}}\left[\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})\right]|z_i(\boldsymbol{s})z_k(\boldsymbol{s})z_l(\boldsymbol{s})|\,\mathrm{d}\boldsymbol{s}.$$

Under conditions (C2) and (C5), $\mathbb{E}_{\boldsymbol{\theta_0}}\{G_{n,ikl}\}$ is a non-negative constant. If it is zero, the product of the covariate terms inside has to be zero almost everywhere on $\mathcal{W}\times\mathcal{T}$, which contradicts condition (C8). Thus $\mathbb{E}_{\boldsymbol{\theta_0}}(G_{n,ikl})$ is strictly positive. Then, by Markov's inequality, for any $\delta>0$, there always exists a finite $H_{ikl}(\delta)=\mathbb{E}_{\boldsymbol{\theta_0}}(G_{n,ikl})/\delta$ such that $\mathbb{P}_{\boldsymbol{\theta_0}}\left[|G_{n,ikl}|\geq H_{ikl}(\delta)\right]\leq\delta$. Since $H_{ikl}(\delta)$ depends only on δ but not on n, $G_{n,ikl}$, and thus (3.19), converges under $P_{\boldsymbol{\theta_0}}$ in probability to $O(1)$. This result applies to every component of (3.19) with $1\leq i,k,l\leq p$.

Move back to the Taylor expansion (3.16). By Lemma 3.4, $-\nabla\boldsymbol{s}_n(\boldsymbol{\theta_0})/n$ converges almost surely, and thus in probability, under $P_{\boldsymbol{\theta_0}}$ to $\boldsymbol{U}$. Collecting all i with $1\leq i\leq p$ and recalling the results obtained above, (3.16) can be rewritten as

$$\left[\boldsymbol{U}+o_P(1)-\frac{1}{2}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})^\top O_P(1)\right](\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})=\frac{\boldsymbol{s}_n(\boldsymbol{\theta_0})}{n}.$$

We appeal to strong consistency (cf., Theorem 3.2) to conclude that $(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})^\top$ times $O_P(1)$ converges under $P_{\boldsymbol{\theta_0}}$ in probability to zero. Furthermore, by condition (C8), the matrix $\boldsymbol{U}$ is invertible. Thus, $\boldsymbol{U}+o_P(1)$ is also invertible with a probability tending to one as $n\to\infty$. Multiplication by this inverse yields that

$$n^{1/2}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})=\left[\boldsymbol{U}+o_P(1)\right]^{-1}\frac{\boldsymbol{s}_n(\boldsymbol{\theta_0})}{n^{1/2}}.$$

The remainder of the proof is a straightforward application of Slutsky's theorem. Obviously, $[\boldsymbol{U}+o_P(1)]^{-1}$ converges under $P_{\boldsymbol{\theta_0}}$ in probability to $\boldsymbol{U}^{-1}$. By Lemma 3.5, $\boldsymbol{s}_n(\boldsymbol{\theta_0})/n^{1/2}$ converges under $P_{\boldsymbol{\theta_0}}$ in distribution to a p-dimensional normally distributed random vector with mean zero and covariance matrix $\boldsymbol{V}$. Thus, as $n\to\infty$, $n^{1/2}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})$ converges under $P_{\boldsymbol{\theta_0}}$ in distribution to a p-dimensional normally distributed random vector with mean zero and covariance matrix $\boldsymbol{U}^{-1}\boldsymbol{V}(\boldsymbol{U}^{-1})^\top$. □

Remark 3.8. *In Theorem 3.2, we have proved that $\widehat{\boldsymbol{\theta}}_n$ is a strongly consistent estimator. Here, observing that $\boldsymbol{U}$ and $\boldsymbol{V}$ are constant matrices, we thus provide the convergence rate: $(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})_l=O_P(n^{-1/2})$ for every $1\leq l\leq p$.*

3.3.3 Existence of the Estimator

Note that both strong consistency and asymptotic normality of the logistic regression estimator rely on the attainment of $\widehat{\boldsymbol{\theta}}_n$. For the sake of completeness, we also formulate the asymptotic existence of the estimator in the next theorem.

Theorem 3.9. *Assume that conditions (C1)–(C8) hold. Define* $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ *by (3.14) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. *Then, for every* n, *an estimator* $\widehat{\boldsymbol{\theta}}_n$ *exists that solves* $\boldsymbol{s}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$ *with a probability tending to one as* $n \to \infty$.

Proof. The proof follows from an appeal to Sørensen (1999, Corollary 2.6). By condition (C2), $\boldsymbol{s}_n(\boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Thus, we only need to verify Sørensen (1999, Condition 2.5). For elegance of writing, we omit the proof here but provide a proof in more general cases in Lemma 3.15. □

3.4 Infill Asymptotics for General Point Process Models

By releasing the Poisson assumption of X_n and modifying the corresponding conditions, the asymptotic results obtained in Section 3.3 can be extended to general models for which replicated point patterns are available. The log-linear form (3.1) for the intensity function may not be always natural though. In this section, we give sufficient conditions for consistency and asymptotic normality and discuss the applicability for some specific families of point processes.

3.4.1 Asymptotic Results

For general point processes, we shall need the following modified conditions.

(C9) Y_i is a general point process with intensity function $\lambda(\boldsymbol{s}; \boldsymbol{\theta})$ given by (3.1), where $b > 0$ is an integrable function, $\boldsymbol{z}$ is a measurable vector of covariates and the parameter vector $\boldsymbol{\theta}$ lies in an open set $\boldsymbol{\Theta} \subset \mathbb{R}^p$; E_i has integrable intensity function $\rho(\boldsymbol{s}) > 0$.

(C10) Both E_i and Y_i have bounded pair correlation functions $g(\boldsymbol{s}, \boldsymbol{u})$ and $h(\boldsymbol{s}, \boldsymbol{u})$: $\sup_{(\boldsymbol{s},\boldsymbol{u})\in(\mathcal{W}\times\mathcal{T})^2} g(\boldsymbol{s}, \boldsymbol{u}) < \infty$ and $\sup_{(\boldsymbol{s},\boldsymbol{u})\in(\mathcal{W}\times\mathcal{T})^2} h(\boldsymbol{s}, \boldsymbol{u}) < \infty$.

The next theorems are concerned with strong consistency, asymptotic normality and existence of logistic regression estimators for general point processes in analogy to Theorems 3.2, 3.7 and 3.9.

Theorem 3.10. *Assume that conditions (C1), (C4)–(C7) and (C9) hold. Define* $L_n(\boldsymbol{\theta}) = L(X_n, D_n; \boldsymbol{\theta})$ *by (3.7) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and set* $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta})$. *If* $\widehat{\boldsymbol{\theta}}_n$ *is attained, as* $n \to \infty$, $\widehat{\boldsymbol{\theta}}_n$ *converges* $P_{\boldsymbol{\theta_0}}$*-almost surely to* $\boldsymbol{\theta_0}$.

Proof. The same proofs as of Lemma 3.1 and Theorem 3.2 can be applied straightforwardly. □

Theorem 3.11. *Assume that conditions (C1), (C4)–(C10) hold. Define* $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ *by (3.14) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and let* $\widehat{\boldsymbol{\theta}}_n$ *be the estimator for which* $\boldsymbol{s}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. *If* $\widehat{\boldsymbol{\theta}}_n$ *is attained, as* $n \to \infty$, $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})$ *converges under* $P_{\boldsymbol{\theta_0}}$ *in distribution to a* p*-dimensional normally distributed random vector with mean zero and covariance matrix* $\boldsymbol{U}^{-1}\boldsymbol{V}(\boldsymbol{U}^{-1})^\top$, *where* $\boldsymbol{U}$ *is as defined in Lemma 3.4 and* $\boldsymbol{V}$ *is now given by*

$$\begin{aligned}\boldsymbol{V} = &\left[\int_{\mathcal{W}\times\mathcal{T}} \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})} z_k(\boldsymbol{s}) z_l(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p} \\ &+ \left[\int\int_{(\mathcal{W}\times\mathcal{T})^2} \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})\rho(\boldsymbol{s})\rho(\boldsymbol{u}) z_k(\boldsymbol{s}) z_l(\boldsymbol{u})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})][\lambda(\boldsymbol{u};\boldsymbol{\theta_0})+\rho(\boldsymbol{u})]} [g(\boldsymbol{s},\boldsymbol{u})-1] \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{u}\right]_{k,l=1}^{p} \\ &+ \left[\int\int_{(\mathcal{W}\times\mathcal{T})^2} \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})\rho(\boldsymbol{s})\rho(\boldsymbol{u}) z_k(\boldsymbol{s}) z_l(\boldsymbol{u})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})][\lambda(\boldsymbol{u};\boldsymbol{\theta_0})+\rho(\boldsymbol{u})]} [h(\boldsymbol{s},\boldsymbol{u})-1] \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{u}\right]_{k,l=1}^{p}.\end{aligned}$$

Proof. Similar proofs as of Lemma 3.4, Lemma 3.5 and Theorem 3.7 can be used. Note that, by condition (C10), the pair correlation function of Y_i, $h(\boldsymbol{s},\boldsymbol{u})$, is bounded, which ensures the entries of $\boldsymbol{V}$ to be finite. □

Theorem 3.12. *Assume that conditions (C1), (C4)–(C10) hold. Define* $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ *by (3.14) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. *Then, for every* n, *an estimator* $\widehat{\boldsymbol{\theta}}_n$ *exists that solves* $\boldsymbol{s}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$ *with a probability tending to one as* $n \to \infty$.

Proof. A similar proof as of Theorem 3.9 can also be applied. □

3.4.2 Applicability Discussion

As an example, consider an inhomogeneous Neyman-Scott point process (Waagepetersen, 2007) with a stationary Poisson parent point process of intensity $\kappa > 0$ and clusters that are independent Poisson point processes with intensity functions $\mu k(\boldsymbol{s}-\boldsymbol{u})\exp[\boldsymbol{\theta}^\top \boldsymbol{z}(\boldsymbol{s})]$ centred around the parent points $\boldsymbol{u}$. Here, μ is the baseline offspring intensity and $k(\cdot)$ is some probability density function determining the spread of offspring points in clusters. The intensity function of the combined offspring process then reads $\kappa\mu\exp[\boldsymbol{\theta}^\top \boldsymbol{z}(\boldsymbol{s})]$, which is of the log linear form (3.1) upon adding a component with the entry one to $\boldsymbol{z}$ and combining κ and μ.

Also the strength of interactions may enter into the intensity function. For instance, consider a log-Gaussian Cox process (Coles & Jones, 1991; Møller, Syversveen & Waagepetersen, 1998) driven by a Gaussian random field with mean function $[\boldsymbol{\theta}^\top \boldsymbol{z}(\boldsymbol{s})]\log[b(\boldsymbol{s})]$ and covariance function $\sigma^2 c(\boldsymbol{s},\boldsymbol{u})$ where $c(\boldsymbol{s},\boldsymbol{u})$ is some correlation function and $\sigma^2 > 0$ denotes the variance at distance zero. Its intensity function reads $\exp[\boldsymbol{\theta}^\top \boldsymbol{z}(\boldsymbol{s}) + \sigma^2/2]$, which still has the log-linear form

(3.1) upon adding a component with the entry one to $\boldsymbol{z}$. Should $c(\boldsymbol{s}, \boldsymbol{u})$ depend on further parameters, e.g. the decay rate of interactions, additional estimating equations are required.

3.5 Construction of Asymptotic Confidence Intervals

Central limit theorems, such as Theorems 3.7 and 3.11, can be used to construct approximate confidence intervals, which enables uncertainty quantification for predictions, for instance, in Chapter 2. To this end, we propose consistent estimators to approximate the asymptotic covariance matrices obtained in Sections 3.3 and 3.4. Afterwards, we conduct a simulation study to demonstrate how a regular dummy process can help reduce the asymptotic variance.

3.5.1 Estimation of the Asymptotic Covariance Matrix

By Theorems 3.7 and 3.11, the entries in the covariance matrix $n^{-1}\boldsymbol{U}^{-1}\boldsymbol{V}(\boldsymbol{U}^{-1})^\top$ of $\widehat{\boldsymbol{\theta}}_n$ decrease as n increases. However, as they depend on the unknown true intensity function $\lambda(\boldsymbol{s}; \boldsymbol{\theta_0})$, they must be estimated. For the Poisson point process model considered in Section 2.4, we used a simple plug-in estimator of $\lambda(\boldsymbol{s}; \boldsymbol{\theta_0})$ based on the dummy point process D_n. Here, we investigate consistent estimators based on replications.

In the context of Poisson point processes X_n (cf., Section 3.3), consider two sequences of independent and identically distributed point processes $A_{1,n} = \cup_{i=1}^n B_{1,i}$ and $A_{2,n} = \cup_{i=1}^n B_{2,i}$ defined on $\mathcal{W} \times \mathcal{T}$ that are mutually independent and also independent from X_n and D_n. Assume that $B_{1,i}$ and $B_{2,i}$ have the same intensity function $\rho(\boldsymbol{s})$ as E_i. Furthermore, let $B_{1,i}$ be a Poisson point process and $B_{2,i}$ distributed as E_i having pair correlation function $g(\boldsymbol{s}, \boldsymbol{u})$. We propose estimators for $\boldsymbol{U}$ and $\boldsymbol{V}$ defined in Theorem 3.7 as follows:

$$n\widehat{\boldsymbol{U}} = \sum_{i=1}^{n} \left[\sum_{\boldsymbol{x} \in B_{1,i}} \frac{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) z_k(\boldsymbol{x}) z_l(\boldsymbol{x})}{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{x})} \right]_{k,l=1}^{p} \tag{3.20}$$

and

$$\begin{aligned} n\widehat{\boldsymbol{V}} = n\widehat{\boldsymbol{U}} \\ + \sum_{i=1}^{n} \left[\sum_{\boldsymbol{x} \in B_{2,i}} \sum_{\boldsymbol{y} \in B_{2,i}, \boldsymbol{y} \neq \boldsymbol{x}} \frac{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{x})} \frac{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{y})} \right]_{k,l=1}^{p} \\ - \sum_{i=1}^{n} \left[\sum_{\boldsymbol{x} \in B_{1,i}} \sum_{\boldsymbol{y} \in B_{1,i}, \boldsymbol{y} \neq \boldsymbol{x}} \frac{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{x})} \frac{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{y})} \right]_{k,l=1}^{p}. \end{aligned} \tag{3.21}$$

Note that $\rho(\boldsymbol{s})$ and $g(\boldsymbol{s},\boldsymbol{u})$ are known explicitly.

Theorem 3.13. *Assume that conditions (C1)–(C8) hold. Define* $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ *by (3.14) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and let* $\widehat{\boldsymbol{\theta}}_n$ *be the estimator for which* $\boldsymbol{s}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. *If* $\widehat{\boldsymbol{\theta}}_n$ *is attained, as* $n \to \infty$, $\widehat{\boldsymbol{U}}$ *and* $\widehat{\boldsymbol{V}}$ *are consistent estimators of the matrices* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *defined in Lemma 3.4 and Lemma 3.5.*

Proof. To show the consistency of approximating $n^{-1}\boldsymbol{U}^{-1}\boldsymbol{V}(\boldsymbol{U}^{-1})^\top$ by $(n\widehat{\boldsymbol{U}})^{-1}n\widehat{\boldsymbol{V}}\ [(n\widehat{\boldsymbol{U}})^{-1}]^\top$ using estimators $n\widehat{\boldsymbol{U}}$ and $n\widehat{\boldsymbol{V}}$, we need to prove that $\widehat{\boldsymbol{U}}$ and $\widehat{\boldsymbol{V}}$ converge in $P_{\boldsymbol{\theta_0}}$-probability to $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively.

For

$$\widehat{\boldsymbol{U}} = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\boldsymbol{x}\in B_{1,i}} \frac{\lambda(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_n) z_k(\boldsymbol{x}) z_l(\boldsymbol{x})}{\lambda(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_n)+\rho(\boldsymbol{x})}\right]_{k,l=1}^{p},$$

we prove component-wise convergence. First, we show that, for every $1 \le k, l \le p$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in B_{1,i}} \frac{\lambda(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_n) z_k(\boldsymbol{x}) z_l(\boldsymbol{x})}{\lambda(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_n)+\rho(\boldsymbol{x})} - \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in B_{1,i}} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0}) z_k(\boldsymbol{x}) z_l(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})+\rho(\boldsymbol{x})}\right| = o_P(1).$$

Since $\lambda(\boldsymbol{x};\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$, one can derive the first-order Taylor series for the terms between the absolute brackets and bound the left-hand side above by

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in B_{1,i}}\left|\frac{\lambda(\boldsymbol{x};\bar{\boldsymbol{\theta}})\rho(\boldsymbol{x}) z_k(\boldsymbol{x}) z_l(\boldsymbol{x})}{\left[\lambda(\boldsymbol{x};\bar{\boldsymbol{\theta}})+\rho(\boldsymbol{x})\right]^2}\sum_{j=1}^{p}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}\right)_j z_j(\boldsymbol{x})\right|,$$

where $\bar{\boldsymbol{\theta}}$ is a convex combination of $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta_0}$. Considering that $0 < \lambda(\boldsymbol{x};\bar{\boldsymbol{\theta}})\rho(\boldsymbol{x})/[\lambda(\boldsymbol{x};\bar{\boldsymbol{\theta}})+\rho(\boldsymbol{x})]^2 < 1$, by condition (C5), one can find a constant vector $\boldsymbol{c} = [c_1,\ldots,c_p]^\top$ such that the term above is further bounded by

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in B_{1,i}}|z_k(\boldsymbol{x})z_l(\boldsymbol{x})|\left|\sum_{j=1}^{p}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}\right)_j c_j\right|. \tag{3.22}$$

Since $\widehat{\boldsymbol{\theta}}_n$ depends on X_n and D_n while $A_{1,n} = \cup_{i=1}^{n} B_{1,i}$ is independent of X_n, D_n, one may separately analyze the two terms: $n^{-1}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in B_{1,i}}|z_k(\boldsymbol{x})\ z_l(\boldsymbol{x})|$ and $|\sum_{j=1}^{p}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})_j c_j|$. Note that the former is the average of independent and identically distributed real-valued random variables. By the Campbell–Mecke theorem,

$$\mathbb{E}_{\boldsymbol{\theta_0}}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in B_{1,i}}|z_k(\boldsymbol{x})z_l(\boldsymbol{x})|\right] = \int_{\mathcal{W}\times\mathcal{T}}|z_k(\boldsymbol{s})z_l(\boldsymbol{s})|\,\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\mathrm{d}\boldsymbol{s}.$$

Condition (C5) implies that the covariate terms in the integrand are bounded. Conditions (C2) and (C5) imply that the intensity function $\lambda(\boldsymbol{s};\boldsymbol{\theta_0})$ is integrable on $\mathcal{W}\times\mathcal{T}$. By condition (C8), the product of the covariate terms cannot be zero almost everywhere on $\mathcal{W}\times\mathcal{T}$. Kolmogorov's strong law of large numbers then implies that $n^{-1}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in B_{1,i}}|z_k(\boldsymbol{x})z_l(\boldsymbol{x})|$ converges $P_{\boldsymbol{\theta_0}}$-almost surely, and thus in $P_{\boldsymbol{\theta_0}}$-probability, to the integral above which is $O(1)$. Under Conditions (C1), (C2) and (C4)–(C7), by Theorem 3.2, $|\sum_{j=1}^{p}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta_0})_j c_j|$ converges in probability to zero. Hence, (3.20) converges in $P_{\boldsymbol{\theta_0}}$-probability to zero.

On the other hand, Kolmogorov's strong law of large numbers also implies that

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in B_{1,i}}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})z_k(\boldsymbol{x})z_l(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})+\rho(\boldsymbol{x})}\overset{P_{\boldsymbol{\theta_0}}-a.s.}{\to}\int_{\mathcal{W}\times\mathcal{T}}\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\rho(\boldsymbol{s})}{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})}z_k(\boldsymbol{s})z_l(\boldsymbol{s})\mathrm{d}\boldsymbol{s},$$

from which convergence in $P_{\boldsymbol{\theta_0}}$-probability follows. By the triangle inequality, $\widehat{\boldsymbol{U}}$ converges in $P_{\boldsymbol{\theta_0}}$-probability to $\boldsymbol{U}$.

The remainder of the proof mostly proceeds along similar lines for the terms in $\widehat{\boldsymbol{V}}$. For instance, we need to show that, for every $1\le k,l\le p$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in B_{2,i}}\sum_{\boldsymbol{y}\in B_{2,i},\boldsymbol{y}\neq\boldsymbol{x}}\frac{\lambda(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_n)z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_n)+\rho(\boldsymbol{x})}\frac{\lambda(\boldsymbol{y};\widehat{\boldsymbol{\theta}}_n)z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y};\widehat{\boldsymbol{\theta}}_n)+\rho(\boldsymbol{y})}\\&\overset{P_{\boldsymbol{\theta_0}}}{\to}\int\int_{(\mathcal{W}\times\mathcal{T})^2}\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})\rho(\boldsymbol{s})\rho(\boldsymbol{u})z_k(\boldsymbol{s})z_l(\boldsymbol{u})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})][\lambda(\boldsymbol{u};\boldsymbol{\theta_0})+\rho(\boldsymbol{u})]}g(\boldsymbol{s},\boldsymbol{u})\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{u}.\end{aligned}$$

To do so, we first note that the absolute value of

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{\boldsymbol{x}\in B_{2,i}}\sum_{\boldsymbol{y}\in B_{2,i},\boldsymbol{y}\neq\boldsymbol{x}}\frac{\lambda(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_n)z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_n)+\rho(\boldsymbol{x})}\frac{\lambda(\boldsymbol{y};\widehat{\boldsymbol{\theta}}_n)z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y};\widehat{\boldsymbol{\theta}}_n)+\rho(\boldsymbol{y})}\right\}-\\&\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{\boldsymbol{x}\in B_{2,i}}\sum_{\boldsymbol{y}\in B_{2,i},\boldsymbol{y}\neq\boldsymbol{x}}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})+\rho(\boldsymbol{x})}\frac{\lambda(\boldsymbol{y};\boldsymbol{\theta_0})z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y};\boldsymbol{\theta_0})+\rho(\boldsymbol{y})}\right\}\end{aligned}$$

is $o_P(1)$. This can be readily seen from arguments similar to those in the proof above. In a second step, we show that

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{\boldsymbol{x}\in B_{2,i}}\sum_{\boldsymbol{y}\in B_{2,i},\boldsymbol{y}\neq\boldsymbol{x}}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})+\rho(\boldsymbol{x})}\frac{\lambda(\boldsymbol{y};\boldsymbol{\theta_0})z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y};\boldsymbol{\theta_0})+\rho(\boldsymbol{y})}\right\}$$

converges in $P_{\boldsymbol{\theta_0}}$-probability to

$$\int\int_{(\mathcal{W}\times\mathcal{T})^2}\frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})\rho(\boldsymbol{s})\rho(\boldsymbol{u})z_k(\boldsymbol{s})z_l(\boldsymbol{u})}{[\lambda(\boldsymbol{s};\boldsymbol{\theta_0})+\rho(\boldsymbol{s})][\lambda(\boldsymbol{u};\boldsymbol{\theta_0})+\rho(\boldsymbol{u})]}g(\boldsymbol{s},\boldsymbol{u})\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{u}$$

by Kolmogorov's strong laws of large numbers, noting that $g(\boldsymbol{s},\boldsymbol{u})$ is bounded by condition (C3) so that the limit is finite. □

For general point processes X_n (cf., Section 3.4), again consider two independent sequences of independent and identically distributed point processes $A_{1,n} = \cup_{i=1}^{n} B_{1,i}$ and $A_{2,n} = \cup_{i=1}^{n} B_{2,i}$ defined as above. Additionally, consider another sequence of independent and identically distributed point processes $A_{3,n} = \cup_{i=1}^{n} B_{3,i}$ where the $B_{3,i}$ are independent and have the same distribution Y_i such that $A_{3,n}$ is independent from $X_n, D_n, A_{1,n}$ and $A_{2,n}$. With $\widehat{\boldsymbol{U}}$ as above, set

$$
\begin{aligned}
n\widehat{\boldsymbol{V}} = n\widehat{\boldsymbol{U}} \\
&+ \sum_{i=1}^{n} \left[\sum_{\boldsymbol{x} \in B_{2,i}} \sum_{\boldsymbol{y} \in B_{2,i}, \boldsymbol{y} \neq \boldsymbol{x}} \frac{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{x})} \frac{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{y})} \right]_{k,l=1}^{p} \\
&+ \sum_{i=1}^{n} \left[\sum_{\boldsymbol{x} \in B_{3,i}} \sum_{\boldsymbol{y} \in B_{3,i}, \boldsymbol{y} \neq \boldsymbol{x}} \frac{\rho(\boldsymbol{x}) z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{x})} \frac{\rho(\boldsymbol{y}) z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{y})} \right]_{k,l=1}^{p} \\
&- 2 \sum_{i=1}^{n} \left[\sum_{\boldsymbol{x} \in B_{1,i}} \sum_{\boldsymbol{y} \in B_{1,i}, \boldsymbol{y} \neq \boldsymbol{x}} \frac{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) z_k(\boldsymbol{x})}{\lambda(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{x})} \frac{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) z_l(\boldsymbol{y})}{\lambda(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) + \rho(\boldsymbol{y})} \right]_{k,l=1}^{p}.
\end{aligned}
\tag{3.23}
$$

The following consistency holds.

Theorem 3.14. *Assume that conditions (C1), (C4)–(C10) hold. Define* $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ *by (3.14) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and let* $\widehat{\boldsymbol{\theta}}_n$ *be the estimator for which* $\boldsymbol{s}_n(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$. *If* $\widehat{\boldsymbol{\theta}}_n$ *is attained, as* $n \to \infty$, $\widehat{\boldsymbol{U}}$ *and* $\widehat{\boldsymbol{V}}$ *are consistent estimators of the matrices* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *defined in Lemma 3.4 and Theorem 3.11.*

Proof. The same proof as of Theorem 3.13 can be applied. Note that, by condition (C10), $h(\boldsymbol{s}, \boldsymbol{u})$ is bounded, which also ensures that

$$
\int \int_{(\mathcal{W} \times \mathcal{T})^2} \frac{\lambda(\boldsymbol{s}; \boldsymbol{\theta_0}) \lambda(\boldsymbol{u}; \boldsymbol{\theta_0}) \rho(\boldsymbol{s}) \rho(\boldsymbol{u}) z_k(\boldsymbol{s}) z_l(\boldsymbol{u})}{[\lambda(\boldsymbol{s}; \boldsymbol{\theta_0}) + \rho(\boldsymbol{s})][\lambda(\boldsymbol{u}; \boldsymbol{\theta_0}) + \rho(\boldsymbol{u})]} h(\boldsymbol{s}, \boldsymbol{u}) \mathrm{d}\boldsymbol{s} \mathrm{d}\boldsymbol{u}
$$

is finite. □

In the literature, bootstrapping and kernel techniques have been proposed to approximate the asymptotic covariance matrix. In contrast to the former (Guan & Loh, 2007), our estimators are consistent and computationally more efficient. An example of the latter is the estimator proposed in Coeurjolly & Guan (2014) but in the increasing-domain regime.

3.5.2 Simulation Study

In this section, we investigate the use of regular dummy point processes with pair correlation functions smaller than one in reducing the variance of logistic regression estimators, referring to Remark 3.6.

For simplicity, we consider two examples of spatial point process models for $X_n = \cup_{i=1}^n Y_i$ on $\mathcal{W} = [0,1]^2$. In the first example, Y_i is an inhomogeneous Poisson point process with intensity function $\lambda(\boldsymbol{s}) = \exp[\theta_1 + \theta_2 w_x + \theta_3 w_y]$. Note that w_x and w_y are spatial coordinates at location $\boldsymbol{s}$. In the second example, Y_i is a log-Gaussian Cox point process driven by a Gaussian random field with mean function $m(\boldsymbol{s}) = \theta_5 w_x + \theta_6 w_y$ and covariance function $c(\boldsymbol{s}, \boldsymbol{u}) = \sigma^2 \exp[-\|\boldsymbol{s} - \boldsymbol{u}\|/\gamma]$. Thus, its intensity function reads $\lambda(\boldsymbol{s}) = \exp[\theta_5 w_x + \theta_6 w_y + \sigma^2/2]$. For elegance, we write $\theta_4 = \sigma^2/2$.

We also consider two options for dummy point processes $D_n = \cup_{i=1}^n E_i$. In the first option, E_i is a homogeneous Poisson point process with intensity function $\rho(\boldsymbol{s}) = \rho_1$ and pair correlation function $g(\boldsymbol{s}, \boldsymbol{u}) \equiv 1$. In the second option, E_i is a Gaussian determinantal point process also with intensity function $\rho(\boldsymbol{s}) = \rho_1$ but with pair correlation function $g(\boldsymbol{s}, \boldsymbol{u}) < 1$. Note that determinantal point processes have also been used for similar efficiency improvement purposes in Monte Carlo integration (Gautier, Bardenet & Valko, 2019). Moreover, their realizations are easy to obtain (Lavancier, Møller & Rubak, 2015), and the pair correlation functions of Gaussian determinantal point processes are known explicitly, $g(\boldsymbol{s}, \boldsymbol{u}) = 1 - \exp[-2(\|\boldsymbol{s} - \boldsymbol{u}\|/\eta)^2]$.

Specifically, we set $\theta_1 = 0$, $\theta_2 = 1$, $\theta_3 = 2$, $\theta_4 = 0.5$, $\theta_5 = 1$, $\theta_6 = 2$, $\gamma = 0.1$, $\rho_1 = 30$ and $\eta = 0.1$. We conduct two experiments, one in which Y_i is Poisson and the other in which Y_i is a log-Gaussian Cox point process. In each experiment, we generate n replicates $Y_1, \ldots Y_n$ and form their union X_n. We also generate n replicates $E_1, \ldots, E_n$ and take the union D_n from either of the two dummy point process options. We carry out logistic regression estimation and report the parameter estimates as well as their approximate variances and 95% confidence intervals using the consistent estimators proposed in Section 3.5.1. The results for $n = 200$, $n = 600$ and $n = 1000$ are shown in Tables 3.1 and 3.2. We observe that, for both point process models, the estimators using the regular dummy point process option achieve a smaller variance than those using the Poisson point process option. Additionally, as expected, such difference in variance decreases as n increases (cf., Theorems 3.7 and 3.11).

3.6 General Unbiased Estimating Equations

The Campbell–Mecke theorem provides the theoretical foundation for a wide range of unbiased estimating equations. Although strong consistency of these estimators may not hold due to the loss of the likelihood interpretation, asymptotic normality and existence of the estimator hold under appropriate conditions. In this section, we extend our central limit theorem (cf., Theorems 3.11 and 3.12) to more general unbiased estimating equations by introducing the necessary background and providing full proofs.

Recalling (3.3), a more general unbiased estimating function based on the

Table 3.1 Simulation results for the Poisson point process model using Poisson or regular dummy point process options in logistic regression estimation.

n	Dummy	Parameter	True	Estimate	Variance	95% CI
200	Poisson	θ_1	0	0.150	0.012	[-0.062, 0.362]
		θ_2	1	0.869	0.014	[0.638, 1.100]
		θ_3	2	1.854	0.016	[1.608, 2.100]
	Regular	θ_1	0	0.131	0.011	[-0.076, 0.338]
		θ_2	1	0.914	0.013	[0.692, 1.135]
		θ_3	2	1.865	0.015	[1.626, 2.104]
600	Poisson	θ_1	0	-0.025	0.0042	[-0.152, 0.102]
		θ_2	1	0.896	0.0048	[0.760, 1.032]
		θ_3	2	2.126	0.0054	[1.982, 2.270]
	Regular	θ_1	0	-0.042	0.0039	[-0.165, 0.081]
		θ_2	1	0.913	0.0046	[0.781, 1.045]
		θ_3	2	2.139	0.0050	[2.000, 2.278]
1000	Poisson	θ_1	0	-0.011	0.0024	[-0.108, 0.086]
		θ_2	1	1.019	0.0028	[0.915, 1.123]
		θ_3	2	2.025	0.0031	[1.915, 2.135]
	Regular	θ_1	0	-0.016	0.0023	[-0.110, 0.078]
		θ_2	1	1.036	0.0026	[0.936, 1.136]
		θ_3	2	2.008	0.0029	[1.902, 2.114]

Table 3.2 Simulation results for the log-Gaussian Cox process model using Poisson or regular dummy point process options in logistic regression estimation.

n	Dummy	Parameter	True	Estimate	Variance	95% CI
200	Poisson	θ_4	0.5	0.507	0.007	[0.342, 0.672]
		θ_5	1	1.053	0.013	[0.832, 1.274]
		θ_6	2	1.907	0.013	[1.684, 2.130]
	Regular	θ_4	0.5	0.511	0.005	[0.372, 0.650]
		θ_5	1	1.052	0.012	[0.838, 1.265]
		θ_6	2	1.902	0.012	[1.687, 2.117]
600	Poisson	θ_4	0.5	0.501	0.0034	[0.386, 0.616]
		θ_5	1	1.049	0.0044	[0.920, 1.178]
		θ_6	2	1.927	0.0045	[1.795, 2.059]
	Regular	θ_4	0.5	0.502	0.0031	[0.393, 0.611]
		θ_5	1	1.014	0.0040	[0.891, 1.137]
		θ_6	2	1.970	0.0042	[1.844, 2.096]
1000	Poisson	θ_4	0.5	0.495	0.0020	[0.408, 0.582]
		θ_5	1	0.981	0.0023	[0.887, 1.075]
		θ_6	2	2.026	0.0027	[1.924, 2.128]
	Regular	θ_4	0.5	0.491	0.0019	[0.407, 0.575]
		θ_5	1	0.996	0.0021	[0.906, 1.086]
		θ_6	2	2.023	0.0025	[1.925, 2.121]

Campbell–Mecke theorem (e.g., Daley & Vere-Jones, 2009) can take the form

$$\boldsymbol{s}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in X} \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) - \sum_{\boldsymbol{x} \in D} \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) \frac{\lambda(\boldsymbol{x}; \boldsymbol{\theta})}{\rho(\boldsymbol{x})} = \boldsymbol{0}, \tag{3.24}$$

for some test vector function $\boldsymbol{f} : \mathcal{W} \times \mathcal{T} \to \mathbb{R}^p$ such that every component of $\boldsymbol{f}$, denoted by f_i with $1 \leq i \leq p$, has the property that $f_i \lambda$ is absolutely integrable. For instance, the test function $\boldsymbol{f}$ may contain some non-negative

quadrature weight functions (e.g., Waagepetersen, 2008; Guan & Shen, 2010) depending on the first- and second-order characteristics of X to reduce the bias caused by numerical approximation or ignoring the interaction structure. In the quasi-likelihood approach (Guan, Jalilian & Waagepetersen, 2015) and the increasing-domain framework, the optimal $\boldsymbol{f}$ is the solution to a Fredholm integral equation.

To extend our central limit theorem to such unbiased estimating equations in the infill asymptotic regime, we shall need the following additional conditions.

(C11) The component functions $f_i(\boldsymbol{s}, \boldsymbol{\theta})$ are twice continuously differentiable with respect to $\boldsymbol{\theta}$. Its components and the first- and second-order partial derivatives, denoted by $\frac{\partial}{\partial\theta_l} f_i(\boldsymbol{s};\boldsymbol{\theta})$ and $\frac{\partial^2}{\partial\theta_k\partial\theta_l} f_i(\boldsymbol{s};\boldsymbol{\theta})$, in a neighbourhood of $\boldsymbol{\theta_0}$ are bounded in absolute value by some functions $d_i(\boldsymbol{s}), d_{il}(\boldsymbol{s})$ and $d_{ikl}(\boldsymbol{s})$, respectively, that are absolutely integrable with respect to $\lambda(\boldsymbol{s};\boldsymbol{\theta_0})$.

(C12) The second-order partial derivatives satisfy an adapted Hölder condition in a neighbourhood of $\boldsymbol{\theta_0}$: there exist some $\alpha \in (0, 1]$ and some functions $e_{ikl}(\boldsymbol{s})$ which are absolutely integrable with respect to $\lambda(u;\boldsymbol{\theta_0})$ such that

$$\left| \frac{\partial^2}{\partial\theta_k\partial\theta_l} f_i(\boldsymbol{s};\boldsymbol{\theta}) - \frac{\partial^2}{\partial\theta_k\partial\theta_l} f_i(\boldsymbol{s};\boldsymbol{\theta_0}) \right| \leq e_{ikl}(\boldsymbol{s}) \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^{\alpha}.$$

(C13) The $[p \times p]$-dimensional matrix $\widetilde{\boldsymbol{U}}$, whose (k,l)-th entry reads $\int_{\mathcal{W}\times\mathcal{T}} f_k(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) z_l(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}$ with $\boldsymbol{\theta_0} \in \boldsymbol{\Theta}$, is invertible.

(C14) The product of any two component functions, $f_k(\boldsymbol{s};\boldsymbol{\theta_0}) f_l(\boldsymbol{s};\boldsymbol{\theta_0})$, is absolutely integrable with respect to $\lambda(\boldsymbol{s};\boldsymbol{\theta_0})$.

A few remarks on the conditions are appropriate. Conditions (C11) and (C12) restrict the smoothness of $\boldsymbol{f}$ and its first-order and second-order derivatives. Condition (C13) is the counterpart of condition (C8) for general unbiased estimating equations. The technical constraints in condition (C14) ensure that the entries in the covariance matrices of the corresponding estimators are finite.

For the elegance of writing, we start our investigations with the following lemma.

Lemma 3.15. *Assume that conditions (C1), (C5), (C9) and (C11) hold. Define* $\boldsymbol{s}_n(\boldsymbol{\theta}) = \boldsymbol{s}(X_n, D_n; \boldsymbol{\theta})$ *by (3.24) with* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. *Then, as* $n \to \infty$, *for all* $\beta > 0$,

$$\sup_{\boldsymbol{\theta'_k} \in C_{n,\beta}} \left\| -\frac{1}{n} \nabla \boldsymbol{s}_n(\boldsymbol{\theta'_1}, \ldots, \boldsymbol{\theta'_p}) - \widetilde{\boldsymbol{U}} \right\|$$

converges under $P_{\boldsymbol{\theta_0}}$ *in probability to zero, where* $C_{n,\beta} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\| \leq \beta/n^{1/2}\}$ *and*

$$\widetilde{\boldsymbol{U}} = \left[\int_{\mathcal{W}\times\mathcal{T}} f_k(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) z_l(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} \right]_{k,l=1}^{p}.$$

Proof. Under conditions (C1), (C9) and (C11), recalling the unbiased estimating equation (3.24), $-\nabla \boldsymbol{s}_n(\boldsymbol{\theta'_1}, \ldots, \boldsymbol{\theta'_p})/n$ reads

$$\frac{1}{n}\left[\sum_{\boldsymbol{x}\in X_n} -\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta'_k})}{\partial\theta_l} + \sum_{\boldsymbol{x}\in D_n}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta'_k})}{\rho(\boldsymbol{x})}\left[\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta'_k})}{\partial\theta_l} + f_k(\boldsymbol{x};\boldsymbol{\theta'_k})z_l(\boldsymbol{x})\right]\right]_{k,l=1}^{p}.$$

Fix $\beta > 0$. We prove the convergence of the supremum component-wise.

First, by condition (C11), for large enough n and every $1 \le i,k,l \le p$,

$$|f_i(\boldsymbol{s};\boldsymbol{\theta})| \le d_i(\boldsymbol{s}), \quad \left|\frac{\partial f_i(\boldsymbol{s};\boldsymbol{\theta})}{\partial\theta_l}\right| \le d_{il}(\boldsymbol{s}), \quad \left|\frac{\partial^2 f_i(\boldsymbol{s};\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}\right| \le d_{ikl}(\boldsymbol{s}), \tag{3.25}$$

for all $\boldsymbol{s} \in \mathcal{W}\times\mathcal{T}$ and $\boldsymbol{\theta} \in C_{n,\beta}$. Note that, taking n large enough, $C_{n,\beta}$ lies entirely within $\boldsymbol{\Theta}$, as the parameter space $\boldsymbol{\Theta}$ is open by condition (C9). Furthermore, under condition (C9), the first-order Taylor expansion of $\lambda(u;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta_0}$ reads

$$\lambda(\boldsymbol{s};\boldsymbol{\theta}) - \lambda(\boldsymbol{s};\boldsymbol{\theta_0}) = b(\boldsymbol{s})\exp\left[\underline{\boldsymbol{\theta}}^\top \boldsymbol{z}(\boldsymbol{s})\right]\sum_{i=1}^{p}(\boldsymbol{\theta}-\boldsymbol{\theta_0})_i z_i(\boldsymbol{s}),$$

where $\underline{\boldsymbol{\theta}}$ is a convex combination of $\boldsymbol{\theta}$ and $\boldsymbol{\theta_0}$. By condition (C5), the covariate terms in the inner product $\underline{\boldsymbol{\theta}}^\top \boldsymbol{z}(\boldsymbol{s})$ and the sum above are bounded. Thus, recalling the definition of $C_{n,\beta}$, when $n\to\infty$, one obtains that $\lambda(\boldsymbol{s};\boldsymbol{\theta}) = \lambda(\boldsymbol{s};\boldsymbol{\theta_0}) + b(\boldsymbol{s})o(1)$ for all $\boldsymbol{s}\in\mathcal{W}\times\mathcal{T}$ and $\boldsymbol{\theta}\in C_{n,\beta}$.

Consider $-\nabla \boldsymbol{s}_n(\boldsymbol{\theta'_1}, \ldots, \boldsymbol{\theta'_p})/n$. In the remainder of this proof, denote its (k,l)-th entry by $[-\nabla \boldsymbol{s}_{n,k}(\boldsymbol{\theta'_k})/n]_l$. The difference between $[-\nabla \boldsymbol{s}_{n,k}(\boldsymbol{\theta'_k})/n]_l$ and $[-\nabla \boldsymbol{s}_n(\boldsymbol{\theta_0})/n]_{k,l}$ reads

$$\begin{aligned}
&-\frac{1}{n}\sum_{\boldsymbol{x}\in X_n}\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta'_k})}{\partial\theta_l} + \frac{1}{n}\sum_{\boldsymbol{x}\in X_n}\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta_0})}{\partial\theta_l}\\
&+\frac{1}{n}\sum_{\boldsymbol{x}\in D_n}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta'_k})}{\rho(\boldsymbol{x})}\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta'_k})}{\partial\theta_l} - \frac{1}{n}\sum_{\boldsymbol{x}\in D_n}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta_0})}{\partial\theta_l}\\
&+\frac{1}{n}\sum_{\boldsymbol{x}\in D_n}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta'_k})}{\rho(\boldsymbol{x})}f_k(\boldsymbol{x};\boldsymbol{\theta'_k})z_l(\boldsymbol{x}) - \frac{1}{n}\sum_{\boldsymbol{x}\in D_n}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}f_k(\boldsymbol{x};\boldsymbol{\theta_0})z_l(\boldsymbol{x}).
\end{aligned} \tag{3.26}$$

Below, we analyze the asymptotic behavior of the second line in (3.26). The analysis of the other two lines proceeds along similar lines.

Under condition (C11), the first-order Taylor expansion of $\frac{\partial}{\partial\theta_l}f_k(\boldsymbol{x};\boldsymbol{\theta'_k})$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta_0}$ reads

$$\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta'_k})}{\partial\theta_l} - \frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta_0})}{\partial\theta_l} = \sum_{i=1}^{p}(\boldsymbol{\theta'_k}-\boldsymbol{\theta_0})_i\frac{\partial^2 f_k(\boldsymbol{x};\boldsymbol{\theta''_k})}{\partial\theta_i\partial\theta_l},$$

where $\boldsymbol{\theta}_{\boldsymbol{k}}''$ is a convex combination of $\boldsymbol{\theta}_{\boldsymbol{k}}'$ and $\boldsymbol{\theta_0}$, and thus $\boldsymbol{\theta}_{\boldsymbol{k}}'' \in C_{n,\beta}$. Recalling that $\lambda(\boldsymbol{x};\boldsymbol{\theta}_{\boldsymbol{k}}') = \lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + b(\boldsymbol{x})o(1)$, the second line in (3.26) is bounded in absolute value by

$$\sum_{\boldsymbol{x}\in D_n}\left[\sum_{i=1}^{p}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{n\rho(\boldsymbol{x})}\left|(\boldsymbol{\theta}_{\boldsymbol{k}}'-\boldsymbol{\theta_0})_i\right|\left|\frac{\partial^2 f_k(\boldsymbol{x};\boldsymbol{\theta}_{\boldsymbol{k}}'')}{\partial\theta_i\partial\theta_l}\right|+\left|\frac{b(\boldsymbol{x})o(1)}{n\rho(\boldsymbol{x})}\right|\left|\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta}_{\boldsymbol{k}}')}{\partial\theta_l}\right|\right],$$

which, recalling the definition of $C_{n,\beta}$ and the bound functions in (3.21), is further bounded by

$$\sum_{i=1}^{p}\left[\sum_{\boldsymbol{x}\in D_n}\frac{\beta}{n^{3/2}}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}d_{kil}(\boldsymbol{x})\right]+\sum_{\boldsymbol{x}\in D_n}\frac{1}{n}\frac{b(\boldsymbol{x})}{\rho(\boldsymbol{x})}|o(1)|d_{kl}(\boldsymbol{x}) \tag{3.27}$$

for large enough n. Note that this bound depends only on n and not on $\boldsymbol{\theta}_{\boldsymbol{k}}'$. Furthermore, by condition (C11), one can apply the Campbell-Mecke theorem and Markov's inequality as in the proof of Theorem 3.7 to obtain that the inner sum in the first term of (3.27) converges under $P_{\boldsymbol{\theta_0}}$ in probability to zero. Since this result applies to every component of the inner sum with $1 \le i \le p$ and also to the second term, (3.27) converges under $P_{\boldsymbol{\theta_0}}$ in probability to zero. Hence,

$$\sup_{\boldsymbol{\theta}_{\boldsymbol{k}}'\in C_{n,\beta}}\left|\frac{1}{n}\sum_{\boldsymbol{x}\in D_n}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta}_{\boldsymbol{k}}')}{\rho(\boldsymbol{x})}\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta}_{\boldsymbol{k}}')}{\partial\theta_l}-\frac{1}{n}\sum_{\boldsymbol{x}\in D_n}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta_0})}{\partial\theta_l}\right| = o_P(1).$$

Applying similar proofs to the other two lines in (3.26) and combining the obtained results, by the triangle inequality, one obtains that

$$\sup_{\boldsymbol{\theta}_{\boldsymbol{k}}'\in C_{n,\beta}}\left|\left[-\frac{1}{n}\nabla\boldsymbol{s}_{n,k}(\boldsymbol{\theta}_{\boldsymbol{k}}')\right]_l-\left[-\frac{1}{n}\nabla\boldsymbol{s}_n(\boldsymbol{\theta_0})\right]_{k,l}\right| = o_P(1). \tag{3.28}$$

Finally, consider the analogue of Lemma 3.4 for the unbiased estimating equation (3.24). Under conditions (C1), (C9) and (C11), $-\nabla\boldsymbol{s}_n(\boldsymbol{\theta_0})/n$ reads

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\boldsymbol{x}\in Y_i}-\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta_0})}{\partial\theta_l}+\sum_{\boldsymbol{x}\in E_i}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}\left[\frac{\partial f_k(\boldsymbol{x};\boldsymbol{\theta_0})}{\partial\theta_l}+f_k(\boldsymbol{x};\boldsymbol{\theta_0})z_l(\boldsymbol{x})\right]\right]_{k,l=1}^{p}.$$

Again, by conditions (C5) and (C11), one can apply the Campbell–Mecke theorem and Kolmogorov's strong law of large numbers to obtain that $-\nabla\boldsymbol{s}_n(\boldsymbol{\theta_0})/n$ converges $P_{\boldsymbol{\theta_0}}$-almost surely, and thus in $P_{\boldsymbol{\theta_0}}$-probability, to $\widetilde{\boldsymbol{U}}$. Accordingly,

$$\left|\left[-\frac{1}{n}\nabla\boldsymbol{s}_n(\boldsymbol{\theta_0})\right]_{k,l}-\widetilde{\boldsymbol{U}}_{k,l}\right| = o_P(1). \tag{3.29}$$

The proof is completed by applying the triangle inequality to (3.28) and (3.29). □

The next theorem is concerned with the extended central limit theorem.

Theorem 3.16. *Assume that conditions (C1), (C4)–(C5) and (C9)–(C14) hold. Define $\boldsymbol{s}_n(\boldsymbol{\theta})$ by (3.24) with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then, as $n \to \infty$, for every n, an estimator $\widehat{\boldsymbol{\theta}}_n$ exists that solves $\boldsymbol{s}_n(\boldsymbol{\theta}) = \mathbf{0}$ with a probability tending to one. Moreover, $\widehat{\boldsymbol{\theta}}_n$ converges under $P_{\boldsymbol{\theta_0}}$ in probability to $\boldsymbol{\theta_0}$, and $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0})$ converges under $P_{\boldsymbol{\theta_0}}$ in distribution to a p-dimensional normally distributed random vector with mean zero and covariance matrix $\widetilde{\boldsymbol{U}}^{-1}\widetilde{\boldsymbol{V}}(\widetilde{\boldsymbol{U}}^{-1})^\top$, where $\widetilde{\boldsymbol{U}}$ is as defined in Lemma 3.15 and $\widetilde{\boldsymbol{V}}$ is given by*

$$\begin{aligned}\widetilde{\boldsymbol{V}} = &\left[\int_{\mathcal{W}\times\mathcal{T}} f_k(\boldsymbol{s};\boldsymbol{\theta_0}) f_l(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{s};\boldsymbol{\theta_0})\left[1 + \frac{\lambda(\boldsymbol{s};\boldsymbol{\theta_0})}{\rho(\boldsymbol{s})}\right]\mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p} \\ &+ \left[\int\int_{(\mathcal{W}\times\mathcal{T})^2} f_k(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) f_l(\boldsymbol{u};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})[g(\boldsymbol{s},\boldsymbol{u}) - 1]\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{u}\right]_{k,l=1}^{p} \\ &+ \left[\int\int_{(\mathcal{W}\times\mathcal{T})^2} f_k(\boldsymbol{s};\boldsymbol{\theta_0})\lambda(\boldsymbol{s};\boldsymbol{\theta_0}) f_l(\boldsymbol{u};\boldsymbol{\theta_0})\lambda(\boldsymbol{u};\boldsymbol{\theta_0})[h(\boldsymbol{s},\boldsymbol{u}) - 1]\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{u}\right]_{k,l=1}^{p}.\end{aligned}$$

Proof. First of all, by condition (C11), $\boldsymbol{s}_n(\boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$. Then, we verify the analogues of Lemmas 3.4 and 3.5 for the unbiased estimating equation (3.24).

The analogue of Lemma 3.4 has been established in Lemma 3.15. For Lemma 3.5, under conditions (C1) and (C9), $\boldsymbol{s}_n(\boldsymbol{\theta})/n^{1/2}$ now reads

$$\frac{\boldsymbol{s}_n(\boldsymbol{\theta_0})}{n^{1/2}} = n^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in Y_i}\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta_0}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in E_i}\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta_0})\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}\right].$$

The two terms between the curly bracket above are both averages of independent and identically distributed real-valued random vectors. By conditions (C9)–(C11), their means are finite and identical. Additionally by conditions (C4) and (C14), their variances are also finite and given by the entries of $\widetilde{\boldsymbol{V}}$. Then, similar to the proof of Lemma 3.5, one can apply the Campbell–Mecke theorem and multi-variate Lindeberg–Lévy's central limit theorem to obtain that $\boldsymbol{s}_n(\boldsymbol{\theta_0})/n^{1/2}$ converges under $P_{\boldsymbol{\theta_0}}$ in distribution to a p-dimensional normally distributed random vector with mean zero and covariance matrix $\widetilde{\boldsymbol{V}}$.

From Sørensen (1999, Condition 2.7, Corollary 2.8 and Theorem 2.9), by condition (C13), we thus only need to prove that

$$\sup_{\boldsymbol{\theta}\in C_{n,\beta}}\left\|-\frac{1}{n}\boldsymbol{s}_n(\boldsymbol{\theta})\right\|, \quad \sup_{\boldsymbol{\theta'_k}\in C_{n,\beta}}\left\|-\frac{1}{n}\nabla\boldsymbol{s}_n(\boldsymbol{\theta'_1},\ldots,\boldsymbol{\theta'_p}) - \widetilde{\boldsymbol{U}}\right\|,$$

and

$$\sup_{\boldsymbol{\theta'_k}\in C_{n,\beta}}\left\|-\frac{1}{n}\nabla^2\boldsymbol{s}_{n,i}(\boldsymbol{\theta'_1},\ldots,\boldsymbol{\theta'_p}) - \widetilde{\boldsymbol{Q}}_i\right\|$$

all converge under $P_{\boldsymbol{\theta_0}}$ in probability to zero, where $\widetilde{\boldsymbol{Q}}_i$ is given by

$$\widetilde{\boldsymbol{Q}}_i = \left[\int_{\mathcal{W}\times\mathcal{T}} f_i(\boldsymbol{s},\boldsymbol{\theta_0})\lambda(\boldsymbol{s},\boldsymbol{\theta_0})z_k(\boldsymbol{s})z_l(\boldsymbol{s})\mathrm{d}\boldsymbol{s}\right]_{k,l=1}^{p} + \left[\int_{\mathcal{W}\times\mathcal{T}} \left[\frac{\partial f_i(\boldsymbol{s},\boldsymbol{\theta_0})}{\partial\theta_k}z_l(\boldsymbol{s}) + \frac{\partial f_i(\boldsymbol{s},\boldsymbol{\theta_0})}{\partial\theta_l}z_k(\boldsymbol{s})\right]\lambda(\boldsymbol{s},\boldsymbol{\theta_0})d\boldsymbol{s}\right]_{k,l=1}^{p}.$$

In Lemma 3.15, we already proved that

$$\sup_{\boldsymbol{\theta'_k}\in C_{n,\beta}} \left\| -\frac{1}{n}\nabla \boldsymbol{s}_n(\boldsymbol{\theta'_1},\ldots,\boldsymbol{\theta'_p}) - \widetilde{\boldsymbol{U}}\right\| \overset{P_{\boldsymbol{\theta_0}}}{\to} \mathbf{0}.$$

The remainder of the proof mostly proceeds along similar lines for the other two cases.

Condition (C12) is required in proving the convergence of some terms in the entries of $-\nabla^2 \boldsymbol{s}_{n,i}(\boldsymbol{\theta'_1},\ldots,\boldsymbol{\theta'_p})/n$. For instance, we need to prove that

$$\sup_{\boldsymbol{\theta'_k}\in C_{n,\beta}} \left| \frac{1}{n}\sum_{\boldsymbol{x}\in D_n} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta'_k})}{\rho(\boldsymbol{x})}\frac{\partial^2 f_k(\boldsymbol{x};\boldsymbol{\theta'_k})}{\partial\theta_j\partial\theta_l} - \frac{1}{n}\sum_{\boldsymbol{x}\in D_n} \frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}\frac{\partial^2 f_k(\boldsymbol{x};\boldsymbol{\theta_0})}{\partial\theta_j\partial\theta_l}\right| = o_P(1).$$

Since $f_k(\boldsymbol{s},\boldsymbol{\theta})$ is only twice continuously differentiable with respect to $\boldsymbol{\theta}$, the proof in Lemma 3.15 does not apply here. However, recalling from the proof of Lemma 3.15 that $\lambda(\boldsymbol{x};\boldsymbol{\theta'_k}) = \lambda(\boldsymbol{x};\boldsymbol{\theta_0}) + b(\boldsymbol{x})o(1)$, the absolute value term above is bounded by

$$\frac{1}{n}\sum_{\boldsymbol{x}\in D_n}\left[\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}\left|\frac{\partial^2 f_k(\boldsymbol{x};\boldsymbol{\theta'_k})}{\partial\theta_j\partial\theta_l} - \frac{\partial^2 f_k(\boldsymbol{x};\boldsymbol{\theta_0})}{\partial\theta_j\partial\theta_l}\right| + \left|\frac{b(\boldsymbol{x})o(1)}{\rho(\boldsymbol{x})}\right|\left|\frac{\partial^2 f_k(\boldsymbol{x};\boldsymbol{\theta'_k})}{\partial\theta_j\partial\theta_l}\right|\right],$$

which, under the adapted Hölder condition in condition (C12), is further bounded by

$$\sum_{\boldsymbol{x}\in D_n}\left[\frac{\beta}{n^{\alpha/2+1}}\frac{\lambda(\boldsymbol{x};\boldsymbol{\theta_0})}{\rho(\boldsymbol{x})}e_{kil}(\boldsymbol{x}) + \frac{1}{n}\frac{b(\boldsymbol{x})}{\rho(\boldsymbol{x})}\left|o(1)\right|d_{kil}(\boldsymbol{x})\right]$$

for large enough n. This bound depends only on n and not on $\boldsymbol{\theta'}_k$. By conditions (C11) and (C12), one can again apply the Campbell-Mecke theorem and Markov's inequality as in the proof of Theorem 3.7 to obtain that the bound converges under $P_{\boldsymbol{\theta_0}}$ in probability to zero. □

Remark 3.17. *Looking back on logistic regression estimators given by (3.4), its component functions are smooth with respect to $\boldsymbol{\theta}$ and thus the second-order derivatives satisfy the adapted Hölder condition. Moreover, its component functions and the first- and second-order derivatives are bounded under condition (C5).*

3.7 Summary

In this chapter, under conditions that are easy to verify in practice, we established strong consistency and asymptotic normality for logistic regression estimators for the parameters in the intensity function of a spatio-temporal point process model under the infill framework. Such a regime is quite natural for Poisson point processes observed in a fixed window, as is the case in our study on chimney fire prediction in Chapter 2 which motivated this work. It is also suitable for general point processes when replications are available. Since, in Chapter 2, the data was collected spatio-temporally, we chose to work in space and time, but the asymptotic results are equally valid in $\mathbb{R}^d$. We also extended our central limit theorem to the estimators solving general unbiased estimating functions that are based on the Campbell–Mecke theorem.

From a practical perspective, in a simulation study, we demonstrated how the results can be used to obtain approximate confidence intervals based on consistent estimators for the asymptotic covariance matrices. Logistic regression estimators are computationally fast, as they can be implemented using standard software. In addition, we showed another advantage that regular dummy point processes can be used to increase precision.

For future work, note that additional higher order estimating equations are usually required to estimate interaction parameters for general point processes. Thus, the first direction would be to study infill asymptotics there as well. Moreover, we assumed that the intensity function is log-linear in covariates, which simplified the computations. We believe, however, that similar asymptotic results can be proved for a much wider class of intensity functions. Finally, it would also be interesting to study infill asymptotic results for Gibbs and Markov point processes under similar appropriate conditions on Papangelou conditional intensity functions.

CHAPTER 4

XGBoostPP: Tree-based Estimation of Point Process Intensity Functions

4.1 Introduction

Intensity estimation is a fundamental task when analyzing a point process, as it defines the likelihood of event occurrences within the observation window. The existing literature explores two primary scenarios: one based on the Cartesian coordinates of events (Diggle, 1985) and the other involving spatial covariates (Guan, 2008; Baddeley et al., 2012). Chapter 2 introduces an important context related to the latter: estimating the intensity function for classic medium-sized point patterns in the presence of a large number of covariates. These data sets typically consist of only hundreds or thousands of events in the observation window and replicates are generally unavailable (e.g., Yin & Sang, 2021; Dong et al., 2023; Zhang et al., 2023). In this chapter, we continue to focus on this research scenario. It is worth noting that, theoretically, modelling the intensity as a function of covariates can achieve consistency (Guan, 2008) and, practically, the intensities at nearby locations may vary due to differences in their covariates whereas distant locations might exhibit similar intensities owing to shared covariates.

Relating spatial covariates to the intensity of a point process has a well-established history. A particularly useful approach is to assume a parametric form for the intensity function, e.g. a log-linear model that formulates the log-intensity as a linear combination of accessible covariates. Extensive studies have been conducted on the estimation accuracy and asymptotic properties of the resulting estimators (Schoenberg, 2005; Waagepetersen & Guan, 2009; Guan, Jalilian & Waagepetersen, 2015; Hessellund et al., 2022; Xu et al., 2023). Although a parametric model is easy to estimate and interpret, it may not be flexible enough for certain applications due to the inherent rigidity of the parametric form.

In contrast, nonparametric approaches impose fewer and less unrealistic assumptions and have gained considerable attention recently. Examples include kernel intensity estimators (Guan, 2008; Baddeley et al., 2012), Gaussian Cox process approaches (Cunningham, Shenoy & Sahani, 2008; Kim, Asami & Toda, 2022) and Bayesian nonparametric models (Yin et al., 2022). Specifically, kernel intensity estimators apply the standard kernel smoother in the covariate space

and are computationally efficient (Baddeley, Rubak & Turner, 2015). However, due to the high local variability and bandwidth selection problems, their estimation accuracy may not be as good as others. Gaussian Cox process approaches model the latent intensity as a function of a Gaussian process and require more computational resources to approximate the intractable integration of the Gaussian process, either via Markov chain Monte Carlo algorithms (Møller & Waagepetersen, 2004) or domain discretization (Rue, Martino & Chopin, 2009; Illian, Sørbye & Rue, 2012). Recent studies have proposed various approximation schemes to reduce the computational demand associated with this integration (e.g., Walder & Bishop, 2017; Aglietti et al., 2019), although the computational cost remains a significant challenge.

Another important limitation of existing nonparametric approaches is the constraint to consider only a small number of covariates, primarily due to the well-known 'curse of dimensionality' phenomenon. For instance, Kernel intensity estimators typically accommodate only one or two spatial covariates without imposing additional, potentially unrealistic restrictions, such as isotropy. Gaussian Cox process approaches, on the other hand, usually treat the intensity as a function of spatial coordinates, thereby lacking the capability to incorporate additional covariates. A noteworthy exception is the augmented permanental process (Kim, Asami & Toda, 2022), which extends the work of Flaxman, Teh & Sejdinovic (2017). In Kim, Asami & Toda (2022), the square root-intensity is assumed to be a Gaussian process defined on a multi-dimensional covariate domain. Although, theoretically, there is no limit on the dimension of the covariate space, our numerical results suggest a significant deterioration in the performance of the augmented permanental process approach as the covariate dimension increases.

With the rapid development of data-collection technologies, an expanding array of covariates has become accessible for point pattern analysis (e.g., Zhu & Xie, 2022). This surge underscores the necessity for an approach adept at handling the dimensionality of covariates in intensity estimation. Neural network-based point processes (Okawa et al., 2019; Zhang, Kong & Zhou, 2023) have been explored to learn the deep dependence of covariates on event occurrences. However, they require large data sets, containing millions of events, to obtain predictive performance and may not be suitable for classic medium-sized point patterns. Tree-based models have proven to excel in predicting the relationship between numerous covariates and the response (Grinsztajn, Oyallon & Varoquaux, 2022) on medium-sized data. Moreover, ensemble methods, such as random forests (Breiman, 2001) and gradient boosting machines (Natekin & Knoll, 2013), leverage the combination of multiple weak learners to build a robust learner, behaving highly effective across various tasks. Surprisingly, to the best of our knowledge, no existing work in the point process literature has employed tree-based models for intensity estimation.

This work fills the gap discussed above. Specifically, our contributions are three-fold: (i) we propose a novel tree-based approach, XGBoostPP, for intensity estimation of classic medium-sized point pattern data, particularly over a high-

dimensional covariate space, and design the associated learning algorithm and validation procedure; (ii) we derive a dynamic weighted loss function, tailored for our tree-based ensemble model whose variable space varies considerably during the fitting process, to improve the estimation efficiency for clustered point patterns; (iii) we demonstrate the superiority of the proposed approach to the state-of-the-art via numerical and real data experiments: XGBoostPP achieves comparable performance to the best baseline, the augmented permanental process approach, when the covariate dimension is small, while it outperforms all existing methods when the dimension increases.

The remainder of this chapter is organized as follows. Section 4.2 introduces the background and literature. Section 4.3 presents the proposed model, along with the design of the loss functions, the learning algorithm, and the procedure for hyperparameter selection. Section 4.4 displays a simulation study to evaluate the performance of our proposed method and compares it against the state-of-the-art approaches. In Section 4.5, we apply the method to real-world data sets to further demonstrate its practical utility. Finally, the chapter concludes with a discussion of the findings, limitations and directions for future research.

4.2 Background and Literature

In this section, we provide a brief overview of the point process models that will be used in the simulation study in Section 4.4, along with a summary of nonparametric intensity estimation approaches documented in the literature.

4.2.1 Point Process Models

Consider a point process X defined on a bounded domain $\mathcal{W} \subset \mathbb{R}^2$ with intensity function $\lambda(\boldsymbol{s})$. Assuming second-order intensity-reweighted stationarity (Baddeley, Møller & Waagepetersen, 2000), the pair correlation function of X, cf. Section 1.1, can be simplified to $g(r)$ at distance $r = \|\boldsymbol{x} - \boldsymbol{y}\|$ where $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{W}$.

Poisson point process

A Poisson point process assumes point independence within X and its log-likelihood function is given by

$$\sum_{\boldsymbol{x} \in X} \log[\lambda(\boldsymbol{x})] - \int_{\mathcal{W}} \lambda(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}. \tag{4.1}$$

Note that, for any Poisson process, the pair correlation function $g(r) \equiv 1$.

Log-Gaussian Cox process

A Cox process is a Poisson process with a random intensity function. In particular, we consider the log-Gaussian Cox process whose log-intensity reads

$\log[\Lambda(\boldsymbol{s})] = \log[\tilde{\lambda}(\boldsymbol{s})] + Y(\boldsymbol{s})$, where $\tilde{\lambda}(\cdot)$ is a deterministic function of covariates and $Y(\cdot)$ is a zero-mean Gaussian process with a covariance kernel $\rho(\cdot, \cdot)$. The kernel can take, but is not limited to, the exponential form

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \tau^2 \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{y}\|}{\sigma}\right) \tag{4.2}$$

with τ^2 as the variance and σ as the scaling parameter. The intensity function of this process is given by $\lambda(\boldsymbol{s}) = \mathbb{E}[\Lambda(\boldsymbol{s})] = \tilde{\lambda}(\boldsymbol{s}) \exp(\tau^2/2)$, and the pair correlation function reads $g(r) = \exp[\tau^2 \exp(-r/\sigma)]$ which quickly approaches one as r grows large.

Neyman-Scott point process

A Neyman-Scott point process is generated in two stages. First, a set of parent points are produced from a homogeneous Poisson process of intensity κ. Then, for each parent $\boldsymbol{a}$, a set of offspring points are generated independently from a Poisson process with an intensity function $\lambda_{\boldsymbol{a}}(\boldsymbol{s}) = k(\boldsymbol{s} - \boldsymbol{a}; \sigma)\tilde{\lambda}(\boldsymbol{s})$ where $k(\cdot; \sigma)$ is a density function parameterized by a scalar σ, quantifying the scattering distance. In particular, the Thomas process assumes $k(\cdot; \sigma)$ to be an isotropic Gaussian kernel with a standard deviation σ. The intensity function of such a process is given by $\lambda(\boldsymbol{s}) = \kappa\tilde{\lambda}(\boldsymbol{s})$, and the pair correlation function reads $g(r) = 1 + \exp(-r^2/4\sigma^2)/(4\pi\kappa\sigma^2)$ which also quickly approaches one as r increases.

4.2.2 Intensity Estimation Approaches

Recent related work on nonparametric point process intensity estimation include kernel intensity estimators, Gaussian Cox process approaches and neural network-based point process models.

Kernel intensity estimators

Kernel intensity estimators (Guan, 2008; Baddeley et al., 2012) are extensions of kernel density estimators (Silverman, 1998) to point pattern data. Similar to kernel density estimators, estimation can be challenging for kernel intensity estimators when the dimension of the covariate space is high (Cronie & Van Lieshout, 2018). To address this challenge, several dimension-reduction tools have been developed (e.g., Guan & Wang, 2010). Implementations of kernel intensity estimators are available in the R-package *spatstat* (Baddeley, Rubak & Turner, 2015), however, only data with at most two covariates is supported.

Gaussian Cox processes

Gaussian Cox process approaches (Cunningham, Shenoy & Sahani, 2008; Adams, Murray & MacKay, 2009; Gunter et al., 2014; Samo & Roberts, 2015; Lloyd et al., 2015; Walder & Bishop, 2017; Donner & Opper, 2018; John & Hensman,

2018; Aglietti et al., 2019; Kim, Asami & Toda, 2022) can be viewed as Bayesian alternatives to the kernel intensity estimators. Gaussian Cox processes model the intensity by a Gaussian process with a positive link function and maximize the posterior probability. Different link functions lead to variants of Gaussian Cox processes, such as log-Gaussian Cox processes (exponential link; Møller, Syversveen & Waagepetersen, 1998), permanental processes (quadratic link; McCullagh & Møller, 2006) and sigmoidal Gaussian Cox processes (sigmoid link; Adams, Murray & MacKay, 2009). To our knowledge, most Gaussian Cox process approaches estimate the intensity as a function of spatial and temporal coordinates, with the exception of the augmented permanental process approach (Kim, Asami & Toda, 2022). It assumes that the square root-intensity is generated from a Gaussian process defined on the covariate domain and tackles the intractable integration via path integral formulation, achieving great computational improvements. However, it is not immune to the curse of dimensionality. In our numerical experiments, a significant deterioration in its performance is observed as the dimension of the covariate space increases.

Neural network-based point processes

Neural network-based point process models (Du et al., 2016; Mei & Eisner, 2017; Zuo et al., 2020; Zhu & Xie, 2022; Okawa et al., 2019; Zhang, Kong & Zhou, 2023) employ deep neural networks to capture the complex occurrence features of point patterns. Related to our work, the deep mixture point process (Okawa et al., 2019) models the intensity as a mixture of kernels centred at finite representative points in the observation window, and the mixture weight of a representative point is estimated by a multi-layer perceptron over its covariates. The deep kernel mixture point process (Zhang, Kong & Zhou, 2023) extends the deep mixture point process by adopting another neural network to project the coordinates of points to a hyper-plane and learns the similarity between the representation point and the estimated point using kernels there. Although, methodologically, both the deep mixture point process and the deep kernel mixture point process can be directly applied for classic medium-sized point patterns, their performance needs to be evaluated since deep structures usually require plenty of data for model training to obtain predictive performance.

4.3 Methodology

In this section, we present the proposed model and the likelihood-based loss functions for nonparametric intensity estimation of classic medium-sized point patterns. We also introduce the additive learning algorithm and the associated hyperparameter selection procedure.

4.3.1 The XGBoostPP Model

Consider a point process X defined on an observation window $\mathcal{W}$. Assume that its intensity λ is a function of a p-dimensional covariate vector $\boldsymbol{z}(\boldsymbol{s})$ with $\boldsymbol{s} \in \mathcal{W}$. To estimate the mapping $\lambda[\boldsymbol{z}(\boldsymbol{s})] : \mathbb{R}^p \to \mathbb{R}^+$, we propose a tree-based ensemble model, XGBoostPP, following the classic scalable tree boosting system (Chen & Guestrin, 2016). The standard XGBoost was designed for general regression or classification tasks and is not directly applicable to point pattern data, especially with specific point correlations. Hence, we propose two carefully designed likelihood-based loss functions, an efficient learning algorithm and a validation procedure to adapt to the point process context.

Formally, our XGBoostPP model estimates the log-intensity of X by K additive trees:

$$\log\{\lambda[\boldsymbol{z}(\boldsymbol{s})]\} = \sum_{k=1}^{K} f_k[\boldsymbol{z}(\boldsymbol{s})], \quad f_k \in \mathcal{F}, \tag{4.3}$$

where $\mathcal{F} = \{f[\boldsymbol{z}(\boldsymbol{s})] = \boldsymbol{\theta}_{q[\boldsymbol{z}(\boldsymbol{s})]}\}$ is the space of regression trees with $q[\boldsymbol{z}(\boldsymbol{s})] : \mathbb{R}^p \to \{1, \ldots, N_q\}$ and $\boldsymbol{\theta}_{q[\boldsymbol{z}(\boldsymbol{s})]} \in \mathbb{R}^{N_q}$. Here, q represents a tree structure that maps the covariate information at a location $\boldsymbol{s}$ to a corresponding leaf index, N_q is the number of leaves in this tree, and $\boldsymbol{\theta}_q$ is the vector of leaf scores. Each tree f_k is uniquely defined by an independent tree structure q_k and the associated leaf score vector $\boldsymbol{\theta}_{q_k}$. We denote the estimated intensity function by $\lambda(\boldsymbol{x}; \boldsymbol{f}_K)$ with $\boldsymbol{f}_K = (f_1, \cdots, f_K)$ and write $\boldsymbol{\theta}_k$ for $\boldsymbol{\theta}_{q_k}$, with $\theta_{k,v}$ representing the element of $\boldsymbol{\theta}_k$ on leaf v. Given a location $\boldsymbol{s}$, the decision rules of the trees in $\boldsymbol{f}_K$ will first classify $\boldsymbol{z}(\boldsymbol{s})$ into K leaves and then calculate the estimated log-intensity by adding up the scores on all corresponding leaves.

4.3.2 Likelihood-based Loss Functions

To estimate the structures of the K tree predictors in the XGBoostPP model, as well as their leaf scores, we design two likelihood-based loss functions: the Poisson likelihood loss and the dynamic weighted likelihood loss.

Poisson likelihood loss

It is well-known that a parametric intensity function defined as $\lambda(\boldsymbol{s}; \boldsymbol{\theta}) = \exp[\boldsymbol{\theta}^\top \boldsymbol{z}(\boldsymbol{s})]$ can be consistently estimated by maximizing the Poisson log-likelihood (4.1), even when the underlying point process is not Poisson (Schoenberg, 2005). Motivated by this phenomenon, we propose to estimate the optimal tree predictors and leaf scores in model (4.3) by minimizing the following penalized loss function

$$L(\boldsymbol{f}_K) = \sum_{k=1}^{K} \Omega(f_k) - \sum_{\boldsymbol{x} \in X} \phi(\boldsymbol{x}; \boldsymbol{f}_K) + \int_{\mathcal{W}} \exp[\phi(\boldsymbol{s}; \boldsymbol{f}_K)] \, \mathrm{d}\boldsymbol{s}, \tag{4.4}$$

where $\phi(\boldsymbol{s}; \boldsymbol{f}_K) = \log[\lambda(\boldsymbol{s}; \boldsymbol{f}_K)]$ and $\Omega(f_k) = \gamma \sum_v |\theta_{k,v}|$ with some $\gamma > 0$. The last two terms in (4.4) compute the negative Poisson log-likelihood and the first term penalizes the complexity of $\boldsymbol{f}_K$ to avoid overfitting.

Note that one needs to minimize (4.4) with respect to both tree structures $\{q_k\}$ and leaf scores $\{\boldsymbol{\theta}_k\}$, simultaneously. However, it is impractical to exhaust all possible tree structures. Hence, for feasible computation, we develop a greedy search algorithm, which will be illustrated in Section 4.3.3, to approximate the optimal intensity estimates.

Dynamic weighted likelihood loss

When the underlying point process deviates from a Poisson process, e.g. the log-Gaussian Cox process and the Neyman-Scott process, maximizing the Poisson log-likelihood, as discussed in Schoenberg (2005), still results in a consistent estimator for a parametric intensity function. However, the estimation efficiency can be poor due to inter-point dependence (Guan & Shen, 2010). Such a limitation also applies to XGBoostPP when minimizing (4.4). To improve the estimation efficiency for clustered processes, we propose a dynamic weighted likelihood loss, inspired by the quasi-likelihood method developed in Guan, Jalilian & Waagepetersen (2015).

To motivate our proposal, we first assume that the optimal tree structures under model (4.3) exist, denote them by $\boldsymbol{q}^{\mathcal{O}}$ and examine the 'oracle' estimator for the associated optimal leaf scores that minimizes $\mathbb{E}[L(\boldsymbol{f}_K)]$ in (4.4). In this case, all values of $\theta_{k,v}$ that do not correspond to a 'terminal' node in $\boldsymbol{q}^{\mathcal{O}}$ are set to zero. Our focus will then shift to the estimation efficiency of the scores on the remaining nodes, denoted by $\boldsymbol{\theta}_{\mathcal{A}}$, with $\mathcal{A}$ indicating the set of the true supporting leaves.

It is important to note that, with the knowledge of the optimal tree structures $\boldsymbol{q}^{\mathcal{O}}$, the XGBoostPP model can be interpreted as a generalized linear model over a vector of transformed covariates $\boldsymbol{Z}_{\mathcal{A}}(\boldsymbol{s})$ that denote the membership identities of a location $\boldsymbol{s}$ with respect to the supporting tree leaves. Specifically, if a point belongs to a leaf in a tree, the corresponding element in the membership vector is assigned one, and zero otherwise. Then, $\lambda(\boldsymbol{s}; \boldsymbol{\theta}_{\mathcal{A}}) = \exp[\boldsymbol{\theta}_{\mathcal{A}}^{\top} \boldsymbol{Z}_{\mathcal{A}}(\boldsymbol{s})]$. In this setting, the estimated intensity function is always positive and differentiable with respect to $\boldsymbol{\theta}_{\mathcal{A}}$.

By the Karush–Kuhn–Tucker optimality condition, the estimate of $\boldsymbol{\theta}_{\mathcal{A}}$ obtained by minimizing (4.4) must satisfy the estimating equation

$$\boldsymbol{e}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_{\mathcal{A}}) = \gamma \, \mathrm{sgn}(\boldsymbol{\theta}_{\mathcal{A}}) + \sum_{\boldsymbol{x} \in X} \frac{\partial \phi(\boldsymbol{x}; \boldsymbol{\theta}_{\mathcal{A}})}{\partial \boldsymbol{\theta}_{\mathcal{A}}} - \int_{\mathcal{W}} \frac{\partial \phi(\boldsymbol{x}; \boldsymbol{\theta}_{\mathcal{A}})}{\partial \boldsymbol{\theta}_{\mathcal{A}}} \lambda(\boldsymbol{s}; \boldsymbol{\theta}_{\mathcal{A}}) \mathrm{d}\boldsymbol{s} = \boldsymbol{0}. \quad (4.5)$$

Here, we change $\phi(\boldsymbol{s}; \boldsymbol{f}_K)$ and $\lambda(\boldsymbol{s}; \boldsymbol{f}_K)$ to $\phi(\boldsymbol{s}; \boldsymbol{\theta}_{\mathcal{A}})$ and $\lambda(\boldsymbol{s}; \boldsymbol{\theta}_{\mathcal{A}})$ due to the knowledge of $\boldsymbol{q}^{\mathcal{O}}$. According to Guan, Jalilian & Waagepetersen (2015), a direct

generalization of $\boldsymbol{e}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_{\mathcal{A}}) = \mathbf{0}$ is to consider the following more general form

$$\boldsymbol{e}_{\boldsymbol{h}}(\boldsymbol{\theta}_{\mathcal{A}}) = \gamma \operatorname{sgn}(\boldsymbol{\theta}_{\mathcal{A}}) + \sum_{\boldsymbol{x}\in X} \boldsymbol{h}(\boldsymbol{x};\boldsymbol{\theta}_{\mathcal{A}}) - \int_{\mathcal{W}} \boldsymbol{h}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\mathrm{d}\boldsymbol{s} = \mathbf{0}, \tag{4.6}$$

where $\boldsymbol{h}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ can be any measurable vector function defined on $\mathcal{W}$ and of the same dimension as $\boldsymbol{\theta}_{\mathcal{A}}$. Following Guan, Jalilian & Waagepetersen (2015), we call (4.6) a first-order penalized estimating function.

To quantify the estimation efficiency of $\boldsymbol{\theta}_{\mathcal{A}}$ for clustered point processes, we may analyze the estimation variance of (4.6). Recall that $\boldsymbol{\theta}_{\mathcal{A}} \neq \mathbf{0}$, and denote the sensitivity matrix

$$-\mathbb{E}\left[\frac{\partial \boldsymbol{e}_{\boldsymbol{h}}(\boldsymbol{\theta}_{\mathcal{A}})}{\partial \boldsymbol{\theta}_{\mathcal{A}}^{\top}}\right] = \int_{\mathcal{W}} \nabla\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\boldsymbol{h}^{\top}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\mathrm{d}\boldsymbol{s} \tag{4.7}$$

by $\boldsymbol{s}_{\boldsymbol{h}}$ and the variance matrix

$$\begin{aligned}\operatorname{Var}\left[\boldsymbol{e}_{\boldsymbol{h}}(\boldsymbol{\theta}_{\mathcal{A}})\right] = &\int_{\mathcal{W}} \boldsymbol{h}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\boldsymbol{h}^{\top}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\mathrm{d}\boldsymbol{s} \\ &+ \int\int_{\mathcal{W}^2} \boldsymbol{h}(\boldsymbol{u};\boldsymbol{\theta}_{\mathcal{A}})\boldsymbol{h}^{\top}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{u};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\left[g(\boldsymbol{s},\boldsymbol{u})-1\right]\mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{s}\end{aligned} \tag{4.8}$$

by $\boldsymbol{\Sigma}_{\boldsymbol{h}}$. Under a proper asymptotic scheme, e.g. as derived in Chapter 3, the estimation covariance matrix reads $\boldsymbol{s}_{\boldsymbol{h}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{h}}(\boldsymbol{s}_{\boldsymbol{h}}^{-1})^{\top}$, and its inverse, $\boldsymbol{G}_{\boldsymbol{h}} = \boldsymbol{s}_{\boldsymbol{h}}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{h}}^{-1}\boldsymbol{s}_{\boldsymbol{h}}$, is known as the Godambe information.

The optimal $\boldsymbol{h}^{*}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ that minimizes the inverse Godambe information reduces the estimation variance for $\boldsymbol{\theta}_{\mathcal{A}}$ the most. By Guan, Jalilian & Waagepetersen (2015), $\boldsymbol{h}^{*}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ must satisfy a Fredholm integral equation of the second kind (Hackbusch, 1995), that is, for any $\boldsymbol{w} \in \mathcal{W}$,

$$\begin{aligned}\int_{\mathcal{W}} \boldsymbol{h}^{*}(\boldsymbol{u};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{u};\boldsymbol{\theta}_{\mathcal{A}})\left[g(\boldsymbol{s}-\boldsymbol{u})-1\right]\mathrm{d}\boldsymbol{u} \\ + \boldsymbol{h}^{*}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}}) = \frac{\partial\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})}{\partial\boldsymbol{\theta}_{\mathcal{A}}}/\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}}).\end{aligned} \tag{4.9}$$

Solving (4.9) for $\boldsymbol{h}^{*}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ requires numerical quadrature approximation (Nyström, 1930) or stochastic approximation (Xu, Waagepetersen & Guan, 2019); both are computationally expensive.

Alternatively, a special case (Guan & Shen, 2010) can be considered:

$$\begin{aligned}\int_{\mathcal{W}} \boldsymbol{h}^{*}(\boldsymbol{u};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{u};\boldsymbol{\theta}_{\mathcal{A}})\left[g(\boldsymbol{s}-\boldsymbol{u})-1\right]\mathrm{d}\boldsymbol{u} \\ \simeq \boldsymbol{h}^{*}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\int_{\mathcal{W}}\left[g(\boldsymbol{s}-\boldsymbol{u})-1\right]\mathrm{d}\boldsymbol{u},\end{aligned} \tag{4.10}$$

assuming that $\boldsymbol{h}^{*}(\boldsymbol{u};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{u};\boldsymbol{\theta}_{\mathcal{A}}) \simeq \boldsymbol{h}^{*}(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ for proximate pairs of locations and that $g(\boldsymbol{s}-\boldsymbol{u}) \simeq 1$ for distant pairs, which yields a closed-form

solution of $\boldsymbol{h}^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ to (4.9):

$$\boldsymbol{h}^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}}) = w^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\frac{\partial\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})}{\partial\boldsymbol{\theta}_{\mathcal{A}}}/\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}}) \tag{4.11}$$

with a location-dependent weight function

$$w^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}}) = \frac{1}{1+\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\int_{\mathcal{W}}[g(\boldsymbol{s}-\boldsymbol{u})-1]\,\mathrm{d}\boldsymbol{u}}. \tag{4.12}$$

Under these assumptions, it is straightforward to show that solving (4.9) with $\boldsymbol{h}^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ is equivalent to minimizing the weighted negative log-likelihood

$$\gamma|\boldsymbol{\theta}_{\mathcal{A}}| - \sum_{\boldsymbol{x}\in X} w^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\log[\lambda(\boldsymbol{x};\boldsymbol{\theta}_{\mathcal{A}})] + \int_{\mathcal{W}} w^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})\mathrm{d}\boldsymbol{s}, \tag{4.13}$$

where $|\cdot|$ indicates the L_1-norm of a vector. However, this weighted negative log-likelihood, in practice, is unobtainable as the optimal tree structures $\boldsymbol{q}^{\mathcal{O}}$ are inaccessible.

For the parametric model with a fixed covariate space as in Guan & Shen (2010), the intensity term in $w^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ needs to be estimated in advance by maximizing the Poisson likelihood and are treated as fixed afterwards, while the integral $\int_{\mathcal{W}}[g(\boldsymbol{s}-\boldsymbol{u})-1]\,\mathrm{d}\boldsymbol{u}$ is approximated by $\int_0^m 2\pi r[g(r)-1]\mathrm{d}r$ for some sufficiently large distance m.

Such an estimation procedure does not suit our XGBoostPP model as the latter has a considerably varying variable space. The intensity estimated by minimizing (4.4) may not properly approximate $\lambda(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$. Instead, we decide to compute the weights dynamically at each iteration of the fitting process and propose the following penalized weighted loss function

$$\begin{aligned} L_w(\boldsymbol{f}_K) = &\sum_{k=1}^{K}\Omega(f_k) - \sum_{\boldsymbol{x}\in X} w(\boldsymbol{x};\boldsymbol{f}_K)\phi(\boldsymbol{x};\boldsymbol{f}_K) \\ &+ \int_{\mathcal{W}} w(\boldsymbol{s};\boldsymbol{f}_K)\exp[\phi(\boldsymbol{s};\boldsymbol{f}_K)]\,\mathrm{d}\boldsymbol{s} \end{aligned} \tag{4.14}$$

to estimate the intensity of X. It is worth emphasizing that the weight function $w(\boldsymbol{s};\boldsymbol{f}_K)$, introduced later in (4.16), depends on $\boldsymbol{f}_K$ and varies based on updated tree structures, which incorporates the approximation of $w^*(\boldsymbol{s};\boldsymbol{\theta}_{\mathcal{A}})$ into the estimation of the optimal $\boldsymbol{f}_K$, simultaneously.

4.3.3 Additive Learning Algorithm

Given the impracticality of exhaustively exploring all possible tree structures in the minimization of (4.4) and (4.14), we adopt a sequential exploration in the space of $\boldsymbol{f}$. To maintain generality, we focus on the minimization of (4.14) in the subsequent demonstration.

Adding regression trees

Let us denote the estimated log-intensity over the $k-1$ trees $\widehat{\boldsymbol{f}}_{k-1} = (f_1, \cdots, f_{k-1})$ by $\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1})$. At the k-th iteration, we add a tree predictor $\widehat{f}_k$ to minimize

$$\begin{aligned} L_w^{(k)}(f_k) = \Omega(f_k) - \sum_{\boldsymbol{x} \in X} w(\boldsymbol{x}; \widehat{\boldsymbol{f}}_k) \left[\widehat{\phi}(\boldsymbol{x}; \widehat{\boldsymbol{f}}_{k-1}) + f_k(\boldsymbol{x}) \right] \\ + \int_{\mathcal{W}} w(\boldsymbol{s}; \widehat{\boldsymbol{f}}_k) \exp \left[\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1}) + f_k(\boldsymbol{s}) \right] \mathrm{d}\boldsymbol{s}. \end{aligned} \tag{4.15}$$

Without the knowledge of f_k and for computational efficiency, we iteratively approximate $w(\boldsymbol{s}; \widehat{\boldsymbol{f}}_k)$ based on $\widehat{\boldsymbol{f}}_{k-1}$:

$$w(\boldsymbol{s}; \widehat{\boldsymbol{f}}_k) \simeq \frac{\omega_k}{1 + \exp \left[\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1}) \right] \left[\widehat{K}(m) - \pi m^2 \right]}. \tag{4.16}$$

The normalizing scalar ω_k here is chosen such that $\int_{\mathcal{W}} w(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1}) \mathrm{d}\boldsymbol{s} = 1$, and $\widehat{K}(m) - \pi m^2$ estimates $\int_0^m 2\pi r[g(r) - 1] \mathrm{d}r$, where $\widehat{K}(\cdot)$ is Ripley's K-function (cf., Section 1.1) and can be computed by the following standard nonparametric estimator (Møller & Waagepetersen, 2004)

$$\widehat{K}(m) = \sum_{\boldsymbol{x} \in X} \sum_{\boldsymbol{y} \in X, \boldsymbol{y} \neq \boldsymbol{x}} \frac{1\{\|\boldsymbol{x} - \boldsymbol{y}\| \leq m\}}{\widehat{\lambda}(\boldsymbol{x}; \widehat{\boldsymbol{f}}) \widehat{\lambda}(\boldsymbol{y}; \widehat{\boldsymbol{f}}) \, |\mathcal{W} \cap \mathcal{W}_{\boldsymbol{x}-\boldsymbol{y}}|}. \tag{4.17}$$

Here, $1\{\cdot\}$ is the indicator function, $\widehat{\lambda}(\cdot; \widehat{\boldsymbol{f}})$ is the estimated intensity that may be replaced with the one fitted using (4.4), and $1/|\mathcal{W} \cap \mathcal{W}_{\boldsymbol{x}-\boldsymbol{y}}|$ is the translation edge correction factor (Ohser & Stoyan, 1981). We manually set $\widehat{K}(m) - \pi m^2$ to zero when it is negative.

By simple calculations, one can apply a quadratic approximation to (4.15) to obtain

$$\begin{aligned} L_w^{(k)}(f_k) \simeq \Omega(f_k) - \sum_{\boldsymbol{x} \in X} \widehat{w}_k(\boldsymbol{x}) \widehat{\phi}(\boldsymbol{x}; \widehat{\boldsymbol{f}}_{k-1}) \\ + \int_{\mathcal{W}} \widehat{w}_k(\boldsymbol{s}) \exp \left[\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1}) \right] \mathrm{d}\boldsymbol{s} - \sum_{\boldsymbol{x} \in X} \widehat{w}_k(\boldsymbol{x}) f_k(\boldsymbol{x}) \\ + \int_{\mathcal{W}} \widehat{w}_k(\boldsymbol{s}) \exp \left[\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1}) \right] \left[f_k(\boldsymbol{s}) + \frac{1}{2} f_k^2(\boldsymbol{s}) \right] \mathrm{d}\boldsymbol{s}. \end{aligned} \tag{4.18}$$

Since $\widehat{w}_k(\boldsymbol{s})$ and $\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1})$ are known at the k-th iteration, one can remove the constant terms in (4.18). The loss to be minimized at this iteration can thus be simplified to

$$\begin{aligned} \tilde{L}_w^{(k)}(f_k) = \Omega(f_k) - \sum_{\boldsymbol{x} \in X} \widehat{w}_k(\boldsymbol{x}) f_k(\boldsymbol{x}) \\ + \int_{\mathcal{W}} \widehat{w}_k(\boldsymbol{s}) \exp \left[\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1}) \right] \left[f_k(\boldsymbol{s}) + \frac{1}{2} f_k^2(\boldsymbol{s}) \right] \mathrm{d}\boldsymbol{s}. \end{aligned} \tag{4.19}$$

To compute the integrals throughout the above derivation, we employ a numerical quadrature approximation. Suppose that the observation window $\mathcal{W}$ can be divided into m grid cells $\mathcal{B}_1, \cdots, \mathcal{B}_m$, and each cell is centered at $\boldsymbol{b}_i$ and has volume $|\mathcal{B}_i|$. Then,

$$\int_{\mathcal{W}} \widehat{w}_k(\boldsymbol{s}) h(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} \simeq \sum_{i=1}^{m} \widehat{w}_k(\boldsymbol{b}_i) h(\boldsymbol{b}_i) |\mathcal{B}_i|, \tag{4.20}$$

where $h(\boldsymbol{s})$ is any measurable function defined on $\mathcal{W}$ such that $\widehat{w}_k(\boldsymbol{s})h(\boldsymbol{s})$ is absolutely integrable.

Growing an added tree

To grow an added tree f_k, we adopt a top-down greedy search algorithm. Starting from the root node, we iteratively identify the optimal split among various candidates until a predefined convergence criterion is reached. The key to developing a computationally feasible algorithm is to efficiently compute the loss reduction upon any node split.

To achieve this goal, we denote the set of locations whose $\boldsymbol{z}(\boldsymbol{s})$ belongs to leaf v by $I_{k,v} = \{\boldsymbol{s} : q_k(\boldsymbol{s}) = v\}$. Then, we can extract the contribution of this leaf to (4.19), expand $\Omega(f_k)$ and replace f_k by $\theta_{k,v}$ to obtain

$$\begin{aligned} \tilde{L}_{wv}^{(k)} &= \gamma|\theta_{k,v}| - \sum_{\boldsymbol{x}\in X} \widehat{w}_k(\boldsymbol{x}) 1\{\boldsymbol{x} \in I_{k,v}\}\theta_{k,v} \\ &\quad + \int_{\mathcal{W}} \widehat{w}_k(\boldsymbol{s}) \exp\left[\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1})\right] 1\{\boldsymbol{s} \in I_{k,v}\} \left(\theta_{k,v} + \frac{1}{2}\theta_{k,v}^2\right) \mathrm{d}\boldsymbol{s}. \end{aligned} \tag{4.21}$$

For any given tree structure q_k, the optimal score $\widehat{\theta}_{k,v}$ on a leaf v, minimizing (4.21), has the following closed-form expression

$$\widehat{\theta}_{k,v} = \frac{\mathrm{sgn}(R_{k,v} - T_{k,v}) \max\left(|R_{k,v} - T_{k,v}| - \gamma, 0\right)}{T_{k,v}}, \tag{4.22}$$

where

$$R_{k,v} = \sum_{\boldsymbol{x}\in X} \widehat{w}_k(\boldsymbol{x}) 1\{\boldsymbol{x} \in I_{k,v}\} \tag{4.23}$$

and

$$T_{k,v} = \int_{\mathcal{W}} \widehat{w}_k(\boldsymbol{s}) \exp[\widehat{\phi}(\boldsymbol{s}; \widehat{\boldsymbol{f}}_{k-1})] 1\{\boldsymbol{s} \in I_{k,v}\} \mathrm{d}\boldsymbol{s}. \tag{4.24}$$

It is important to mention that, when $\widehat{\theta}_{k,v} = 0$, the leaf will be eliminated from the tree. Hence, the L_1-penalized loss always tends to produce a smaller tree.

Consequently, for any q_k, one can compute a performance score by plugging in all optimal leaf scores as above into the loss (4.19). Utilizing these performance scores, one can find the best split that minimizes (4.19) among various candidate splits for each node and grow the tree sequentially. To exhaust possible candidate splits at a node, we use Chen & Guestrin (2016, Algorithm 1).

More stable learning

To achieve a more stable learning performance during the fitting of the XGBoostPP model, we also incorporate the shrinkage and subsampling techniques of the standard XGBoost (Chen & Guestrin, 2016). The former shrinks the impacts of the newly added tree by a parameter $\eta > 0$, which is well-known as the learning rate in stochastic optimization problems. By replacing f_k with ηf_k, one leaves more room for future trees to improve model performance. The latter randomly subsamples covariates as the candidate dependent variables for each node split, leading to a boosted random forest.

4.3.4 Hyperparameter Selection

Selecting proper hyperparameters, including the number of tree predictors K, the learning rate η and the penalty scalar γ, is crucial for a well-performing XGBoostPP. To this end, we propose a two-fold cross-validation method.

The validation procedure can be approximately regarded as model fitting on two thinned point processes, both of which have around a half intensity and the same second-order information of the original process (Cronie, Moradi & Biscio, 2023). Specifically, we randomly split data points into two parts and, at each time, use one part to train the model and compute the Poisson log-likelihood based on the estimated intensities of the other test part. We summarize the Poisson log-likelihoods on two test parts as the performance measure. To further reduce the randomness in the data splitting, we repeat such a procedure three times and report the averaged test Poisson log-likelihood as the final evaluation metric to choose the optimal combination of hyperparameters.

4.4 Numerical Study

To evaluate the performance of XGBoostPP, we conduct a numerical study on synthetic data from the point process models introduced in Section 4.2.1 and test on various covariate complexities in different scenarios. The intensity function of the simulated point pattern is governed by a set of covariates that are independently generated from an isotropic Gaussian process with a covariance function $\exp(-10r)$. For comparison, we adopt kernel intensity estimators (KIE) with the ratio (ra) and the reweight (re) methods (Baddeley et al., 2012). For details of the two methods, please refer to Baddeley, Rubak & Turner (2015). It is again worth noting that kernel intensity estimators implemented in the R-package *spatstat* only support data with at most two covariates. For Gaussian Cox process approaches, we implement the augmented permanental process (APP) with the naive kernel (naive) and the degenerate kernel of the random Fourier map (rfm) (Kim, Asami & Toda, 2022). The naive kernel also encounters computational problems when applied to a high-dimensional covariate space. For neural network-based point process models, we reproduce the deep kernel mixture point

process (DKMPP) with the radial basis function kernel (rbf) (Zhang, Kong & Zhou, 2023).

We use two evaluation metrics to measure the performance of all approaches: the Poisson log-likelihoods on new randomly generated test data under the same model and the integrated absolute errors from true intensities. The former focuses on important locations where points occur more often and the latter reads $\int_{\mathcal{W}} |\lambda(\boldsymbol{s}) - \widehat{\lambda}(\boldsymbol{s})| \mathrm{d}\boldsymbol{s}$ with $\lambda(\boldsymbol{s})$ and $\widehat{\lambda}(\boldsymbol{s})$ denoting the theoretical and estimated intensities. For each method to be tested, we report the averaged values and include the standard deviations of the two metrics across 500 simulations. In experiments, we employ the subscripts 'p' and 'w' to represent the XGBoostPP models fitted using the Poisson likelihood loss and the dynamic weighted likelihood loss, respectively. The hyperparameters – K, γ and η – are chosen from the candidate sets $\{1, \ldots, 600\}$, $\{10, 30, 50\}$ and $\{0.1, 0.05, 0.01\}$. Moreover, we employ ten parallel trees for each training iteration of XGBoostPP and set the proportion of the subsampled covariates at a node split to 1/3.

4.4.1 Low-dimensional Covariate Space

The intensity function in this scenario is determined by two covariates, denoted by $z_1(\boldsymbol{s})$ and $z_2(\boldsymbol{s})$. For the Poisson process, we set $\lambda(\boldsymbol{s}) = \alpha \exp\{\beta[z_1(\boldsymbol{s}) + z_2(\boldsymbol{s})]\}$, where $\beta = 0.5$ or 1.0 and α is chosen such that the expected number of events in $\mathcal{W} = [0, 1]^2$ is 400. For the log-Gaussian Cox process, we simulate data from a Poisson process with the random intensity $\Lambda(\boldsymbol{s}) = \alpha \exp\{\beta[z_1(\boldsymbol{s}) + z_2(\boldsymbol{s})] + Y(\boldsymbol{s})\}$, where $Y(\boldsymbol{s})$ is a zero-mean isotropic Gaussian random field with a covariance function $\tau^2 \exp(-r/\sigma)$ and $\tau^2 = 1.0$ or 2.0 and $\sigma = 0.02$ or 0.04. For the Neyman-Scott process, we generate parent points from a homogeneous Poisson process with intensity $\kappa = 100$ or 200 and, for each parent, simulate a Poisson number of offspring points, $\max_{\boldsymbol{s} \in \mathcal{W}}[\alpha \exp\{\beta[z_1(\boldsymbol{s}) + z_2(\boldsymbol{s})]\}]$, where the location of each relative to its parent follows a Gaussian distribution with standard deviation $\sigma = 0.02$ or 0.04. We then keep a offspring point with a probability $\exp\{\beta[z_1(\boldsymbol{s}) + \beta z_2(\boldsymbol{s})]\} / \max_{\boldsymbol{s} \in \mathcal{W}}[\exp\{\beta[z_1(\boldsymbol{s}) + z_2(\boldsymbol{s})]\}]$. The distance m used to approximate the weights of (4.16) is set to 0.06 for the Poisson process and to 3σ for the other two processes.

We report the Poisson log-likelihood and the integrated absolute error results on the Poisson process test data in Tables 4.1 and 4.2 and those on the log-Gaussian Cox process test data and the Neyman-Scott process test data in Tables 4.3, 4.4, 4.5 and 4.6. In general, the integrated absolute error conveys similar messages to those reflected by the Poisson log-likelihood. Specifically, XGBoostPP achieves competitive performance on all types of data under various parameter settings, although the augmented permanental process mostly exhibits the best behavior. The kernel intensity estimator performs poorly as spatial heterogeneity increases, as it tends to produce overly smooth estimates. The deep kernel mixture point process seems unsuitable for this estimation task, which may be due to the lack of data for sufficient training. For all approaches, intensity estimation is challenging when observed point patterns are more spatially

varying (i.e., large β) and clustered (i.e., large τ^2 and σ for the log-Gaussian Cox process and small κ and σ for the Neyman-Scott process). Moreover, comparing $\text{XGBoostPP}_\text{p}$ with $\text{XGBoostPP}_\text{w}$, it is evident that the performance improves on clustered processes while it remains almost unchanged on Poisson processes. This indicates that the proposed dynamic weighted likelihood loss and the designed learning algorithm work well when the clustering feature of a point pattern is unknown. Furthermore, these improvements are greater when point patterns are more spatially heterogeneous and clustered, aligning with the findings in Guan & Shen (2010).

Table 4.1 Averaged Poisson log-likelihoods (standard deviations) of different intensity estimation approaches on the Poisson process test data.

Covariates	**Poisson**	β**=0.5**	β**=1.0**
2 (§4.4.1)	True	2089.4	2290.3
	KIE_ra	2072.9(3.2)	2226.6(5.9)
	APP_naive	2077.7(3.9)	2273.6(4.3)
	APP_rfm	2077.8(3.9)	2273.5(4.2)
	DKMPP_rbf	2055.2(12.5)	2195.5(46.9)
	$\text{XGBoostPP}_\text{p}$	2073.2(3.0)	2261.4(5.0)
	$\text{XGBoostPP}_\text{w}$	2073.2(3.0)	2261.6(5.3)
10 (§4.4.3)	APP_rfm	1943.8(15.0)	2127.3(14.4)
	DKMPP_rbf	2050.7(5.3)	2180.9(30.8)
	$\text{XGBoostPP}_\text{p}$	2062.3(4.5)	2245.4(5.5)
	$\text{XGBoostPP}_\text{w}$	2062.4(4.5)	2245.4(5.5)
Covariates	**Poisson**	β**=0.2**	β**=0.4**
10 (§4.4.2)	True	2048.1	2160.9
	APP_rfm	1857.9(41.2)	1962.2(18.0)
	DKMPP_rbf	2015.0(4.9)	2088.7(8.9)
	$\text{XGBoostPP}_\text{p}$	2025.9(3.8)	2115.9(6.3)
	$\text{XGBoostPP}_\text{w}$	2025.9(3.8)	2115.9(6.3)

Table 4.2 Averaged integrated absolute errors (standard deviations) of different intensity estimation approaches on the Poisson process test data.

Covariates	**Poisson**	β**=0.5**	β**=1.0**
2 (§4.4.1)	KIE_ra	91.1(10.2)	170.0(8.9)
	APP_naive	59.8(11.0)	75.8(11.1)
	APP_rfm	60.7(10.6)	77.7(10.6)
	DKMPP_rbf	125.6(25.9)	236.1(52.4)
	$\text{XGBoostPP}_\text{p}$	86.0(10.3)	103.9(12.0)
	$\text{XGBoostPP}_\text{w}$	86.1(10.3)	103.7(12.2)
10 (§4.4.3)	APP_rfm	191.7(7.2)	218.4(7.7)
	DKMPP_rbf	134.8(9.8)	222.5(35.0)
	$\text{XGBoostPP}_\text{p}$	116.9(12.0)	137.0(11.3)
	$\text{XGBoostPP}_\text{w}$	116.9(12.0)	137.0(11.3)
Covariates	**Poisson**	β**=0.2**	β**=0.4**
10 (§4.4.2)	APP_rfm	210.0(15.0)	226.5(7.9)
	DKMPP_rbf	124.2(9.4)	178.3(12.3)
	$\text{XGBoostPP}_\text{p}$	109.3(10.2)	149.8(11.9)
	$\text{XGBoostPP}_\text{w}$	109.3(10.2)	149.8(11.8)

Table 4.3 Averaged Poisson log-likelihoods (standard deviations) of different intensity estimation approaches on the log-Gaussian Cox process test data.

Covariates	Log-Gaussian Cox	τ^2=1		τ^2=2	
		σ=0.02	σ=0.04	σ=0.02	σ=0.04
			β=0.5		
2 (§4.4.1)	True	2070.2	2080.5	2089.7	2102.5
	KIE_{ra}	2052.7(5.4)	2060.5(7.8)	2071.5(7.1)	2077.4(12.8)
	$\text{APP}_{\text{naive}}$	2057.2(6.6)	2064.3(9.3)	2074.5(8.2)	2075.7(17.1)
	APP_{rfm}	2057.3(6.6)	2064.6(9.2)	2074.8(8.2)	2076.2(16.7)
	$\text{DKMPP}_{\text{rbf}}$	2029.4(11.3)	2021.0(12.8)	2029.9(11.8)	1995.7(17.2)
	$\text{XGBoostPP}_{\text{p}}$	2051.8(5.0)	2059.0(8.0)	2069.0(6.9)	2074.0(14.7)
	$\text{XGBoostPP}_{\text{w}}$	2052.6(5.7)	2060.9(7.7)	2072.1(6.8)	2078.9(12.9)
10 (§4.4.3)	APP_{rfm}	1909.3(21.7)	1901.5(28.9)	1907.1(28.3)	1872.4(42.0)
	$\text{DKMPP}_{\text{rbf}}$	2022.5(9.6)	2007.7(14.5)	2018.7(13.4)	1973.1(22.4)
	$\text{XGBoostPP}_{\text{p}}$	2042.5(6.5)	2049.4(8.8)	2059.0(11.0)	2054.5(25.0)
	$\text{XGBoostPP}_{\text{w}}$	2044.2(6.9)	2051.6(9.0)	2062.7(8.4)	2061.5(16.3)
			β=1.0		
2 (§4.4.1)	True	2304.8	2269.1	2256.0	2283.2
	KIE_{ra}	2238.8(11.4)	2202.3(15.7)	2188.2(15.1)	2211.1(22.8)
	$\text{APP}_{\text{naive}}$	2284.7(8.2)	2244.0(12.6)	2229.6(13.4)	2243.4(27.7)
	APP_{rfm}	2284.7(8.1)	2244.0(12.5)	2229.9(12.8)	2244.4(25.9)
	$\text{DKMPP}_{\text{rbf}}$	2166.2(45.2)	2126.3(43.6)	2143.1(43.0)	2094.6(43.5)
	$\text{XGBoostPP}_{\text{p}}$	2272.2(8.0)	2231.8(11.8)	2219.4(11.6)	2234.5(21.6)
	$\text{XGBoostPP}_{\text{w}}$	2275.8(8.7)	2235.5(12.9)	2228.8(9.6)	2244.9(18.6)
10 (§4.4.3)	APP_{rfm}	2116.5(22.6)	2062.9(30.7)	2038.5(31.7)	2021.0(51.3)
	$\text{DKMPP}_{\text{rbf}}$	2160.2(30.1)	2120.7(29.8)	2140.0(29.8)	2077.3(31.2)
	$\text{XGBoostPP}_{\text{p}}$	2256.4(8.3)	2214.6(12.6)	2199.3(14.7)	2207.3(32.8)
	$\text{XGBoostPP}_{\text{w}}$	2258.8(9.1)	2215.7(13.0)	2208.5(12.2)	2215.1(23.5)
			β=0.2		
10 (§4.4.2)	True	2032.9	2066.7	2080.8	2080.7
	APP_{rfm}	1848.3(31.6)	1871.8(30.8)	1879.9(34.4)	1838.5(42.0)
	$\text{DKMPP}_{\text{rbf}}$	1983.7(9.1)	1961.7(14.0)	1959.2(12.1)	1903.7(21.5)
	$\text{XGBoostPP}_{\text{p}}$	2011.5(4.3)	2041.7(7.0)	2054.1(11.1)	2040.5(19.0)
	$\text{XGBoostPP}_{\text{w}}$	2012.8(4.2)	2042.4(7.0)	2056.0(7.8)	2042.3(15.0)
			β=0.4		
10 (§4.4.2)	True	2134.6	2181.9	2153.3	2171.0
	APP_{rfm}	1923.8(24.5)	1944.8(30.6)	1916.3(30.3)	1887.6(45.7)
	$\text{DKMPP}_{\text{rbf}}$	2060.9(9.9)	2029.6(15.5)	2015.8(13.4)	1993.6(23.5)
	$\text{XGBoostPP}_{\text{p}}$	2092.0(7.8)	2130.7(12.4)	2106.2(12.4)	2105.4(24.6)
	$\text{XGBoostPP}_{\text{w}}$	2096.0(8.7)	2133.5(11.8)	2112.6(10.8)	2110.6(20.3)

4.4.2 Higher-dimensional Covariate Space

In this scenario, we conduct similar experiments but increase the number of accessible covariates to ten. Denote the covariates by $z_1(\boldsymbol{s}), \ldots, z_{10}(\boldsymbol{s})$. The intensity function is now defined as $\lambda(\boldsymbol{s}) = \alpha \exp(\beta\{z_1(\boldsymbol{s}) + z_2(\boldsymbol{s})z_3(\boldsymbol{s})/2 + \exp[z_4(\boldsymbol{s})]/6 + z_5(\boldsymbol{s})^2/2 + 3\sin[z_6(\boldsymbol{s})]\})$. Note that $z_7(\boldsymbol{s}), \ldots, z_{10}(\boldsymbol{s})$ are included as nuisance variables which exist often in real-world applications. We set σ, τ^2, κ to the same values as in Section 4.4.1 while $\beta = 0.2$ or 0.4 to balance the spatial heterogene-

Table 4.4 Averaged integrated absolute errors (standard deviations) of different intensity estimation approaches on the log-Gaussian Cox process test data.

Covariates	Log-Gaussian Cox	τ^2=1		τ^2=2	
		σ=0.02	σ=0.04	σ=0.02	σ=0.04
		β=0.5			
2 (§4.4.1)	KIE_{ra}	93.7(14.7)	99.5(19.6)	96.8(18.0)	110.8(29.8)
	APP_{naive}	65.9(15.4)	78.0(23.0)	74.9(20.0)	101.7(38.2)
	APP_{rfm}	66.3(15.0)	77.9(22.9)	74.8(19.7)	100.8(37.7)
	$DKMPP_{rbf}$	138.3(23.2)	147.3(23.7)	145.2(21.7)	158.0(25.5)
	$XGBoostPP_p$	94.1(14.9)	102.6(20.3)	100.5(18.4)	120.0(31.2)
	$XGBoostPP_w$	92.6(17.0)	97.7(21.1)	92.1(19.2)	108.4(29.2)
10 (§4.4.3)	APP_{rfm}	200.2(9.9)	213.2(17.5)	213.5(18.0)	243.9(36.0)
	$DKMPP_{rbf}$	147.9(13.5)	165.1(20.4)	159.6(18.5)	184.2(30.8)
	$XGBoostPP_p$	118.8(16.1)	126.9(19.8)	125.8(20.3)	155.0(36.6)
	$XGBoostPP_w$	115.1(16.8)	121.2(19.8)	116.0(17.0)	141.5(26.2)
		β=1.0			
2 (§4.4.1)	KIE_{ra}	171.5(12.3)	175.7(16.6)	174.0(15.9)	182.0(24.4)
	APP_{naive}	84.0(15.4)	98.7(26.8)	97.2(27.0)	125.7(52.2)
	APP_{rfm}	85.3(15.1)	99.2(26.5)	97.6(26.3)	124.7(50.5)
	$DKMPP_{rbf}$	237.0(47.9)	238.8(44.6)	238.0(44.6)	240.7(39.4)
	$XGBoostPP_p$	113.3(17.3)	127.1(25.4)	123.8(23.1)	146.2(40.7)
	$XGBoostPP_w$	108.6(18.2)	121.8(26.8)	110.0(18.9)	131.8(33.6)
10 (§4.4.3)	APP_{rfm}	232.0(12.6)	247.4(23.0)	250.8(25.5)	280.1(50.7)
	$DKMPP_{rbf}$	236.8(32.7)	249.3(33.8)	246.9(32.8)	263.2(35.3)
	$XGBoostPP_p$	146.1(14.8)	162.2(22.2)	164.4(22.3)	193.8(39.8)
	$XGBoostPP_w$	142.2(15.9)	160.2(22.7)	147.5(19.1)	179.1(31.6)
		β=0.2			
10 (§4.4.2)	APP_{rfm}	208.0(12.2)	214.2(14.2)	214.2(14.9)	240.9(32.6)
	$DKMPP_{rbf}$	135.9(13.5)	155.6(21.1)	148.7(18.2)	176.3(31.0)
	$XGBoostPP_p$	109.2(12.0)	116.9(17.7)	117.6(20.6)	142.0(34.3)
	$XGBoostPP_w$	104.9(11.8)	112.9(16.8)	110.5(16.4)	135.6(26.4)
		β=0.4			
10 (§4.4.2)	APP_{rfm}	234.4(10.2)	246.1(17.5)	246.5(17.0)	272.7(40.1)
	$DKMPP_{rbf}$	189.2(14.2)	203.1(21.0)	199.7(18.8)	222.1(29.2)
	$XGBoostPP_p$	149.5(14.3)	160.1(20.0)	156.0(20.5)	180.4(35.8)
	$XGBoostPP_w$	140.9(16.7)	154.3(20.3)	139.7(17.3)	166.9(29.8)

ity to a reasonable scale. Considering that the detection of clustering patterns over complex covariate relationships is difficult, we change m for approximating the weights of (4.16) to 0.04 for the Poisson process and to 2σ for the other two models.

The relevant Poisson log-likelihood results are also reported in Tables 4.1, 4.3 and 4.5, and the integrated absolute error results are displayed in Tables 4.2, 4.4 and 4.6. Again, the two evaluation metrics give similar information. In this scenario, the performance of the augmented permanental process deteriorates considerably while the deep kernel mixture point process outperforms it, showcasing the advantage of deep neural networks in approximating complex reponse-covariate relationships. XGBoostPP achieves the best performance and

Table 4.5 Averaged Poisson log-likelihoods (standard deviations) of different intensity estimation approaches on the Neyman-Scott process test data.

Covariates	Neyman-Scott	κ=100		κ=200	
		σ=0.02	σ=0.04	σ=0.02	σ=0.04
			β=0.5		
2 (§4.4.1)	True	2090.9	2074.8	2091.2	2071.7
	KIE_{ra}	2071.1(8.0)	2055.2(6.9)	2072.2(6.2)	2053.4(5.1)
	$\text{APP}_{\text{naive}}$	2073.3(10.4)	2060.2(8.0)	2077.5(7.2)	2059.0(5.9)
	APP_{rfm}	2073.6(10.2)	2060.4(8.0)	2077.6(7.2)	2059.1(5.8)
	$\text{DKMPP}_{\text{rbf}}$	2043.6(12.4)	2030.1(12.4)	2047.0(11.2)	2031.7(12.1)
	$\text{XGBoostPP}_{\text{p}}$	2068.0(8.0)	2054.6(6.9)	2070.7(6.1)	2052.8(4.8)
	$\text{XGBoostPP}_{\text{w}}$	2071.2(8.1)	2056.2(7.7)	2072.6(6.7)	2053.1(5.0)
10 (§4.4.3)	APP_{rfm}	1890.8(34.2)	1905.0(26.5)	1919.6(24.6)	1912.1(21.0)
	$\text{DKMPP}_{\text{rbf}}$	2022.1(17.8)	2014.2(14.8)	2035.3(13.2)	2021.5(10.6)
	$\text{XGBoostPP}_{\text{p}}$	2057.6(11.5)	2044.4(8.4)	2061.4(7.3)	2043.0(6.2)
	$\text{XGBoostPP}_{\text{w}}$	2060.9(9.8)	2046.4(8.3)	2064.0(7.7)	2043.2(6.3)
			β=1.0		
2 (§4.4.1)	True	2263.1	2255.6	2272.5	2306.9
	KIE_{ra}	1995.3(17.5)	1990.0(14.2)	2206.2(13.5)	2238.9(10.8)
	$\text{APP}_{\text{naive}}$	2233.6(16.0)	2233.5(11.8)	2248.9(11.7)	2285.6(8.1)
	APP_{rfm}	2234.4(15.1)	2233.6(11.6)	2248.9(11.9)	2285.8(8.1)
	$\text{DKMPP}_{\text{rbf}}$	2125.9(44.9)	2127.7(44.6)	2145.6(44.6)	2181.7(47.8)
	$\text{XGBoostPP}_{\text{p}}$	2222.1(13.9)	2219.9(10.8)	2236.4(10.6)	2273.2(7.7)
	$\text{XGBoostPP}_{\text{w}}$	2234.1(12.6)	2223.7(11.2)	2246.2(9.5)	2274.2(8.4)
10 (§4.4.3)	APP_{rfm}	2022.2(39.8)	2054.3(31.8)	2063.8(31.7)	2102.1(24.9)
	$\text{DKMPP}_{\text{rbf}}$	2107.0(31.5)	2117.7(29.3)	2138.8(27.3)	2181.7(32.2)
	$\text{XGBoostPP}_{\text{p}}$	2202.1(17.2)	2203.3(11.8)	2220.9(11.4)	2257.2(8.5)
	$\text{XGBoostPP}_{\text{w}}$	2211.7(19.9)	2204.6(11.9)	2228.7(11.9)	2257.2(8.5)
			β=0.2		
10 (§4.4.2)	True	1977.8	1963.7	2070.6	2047.9
	APP_{rfm}	1777.1(33.2)	1786.8(27.5)	1884.5(27.5)	1865.6(29.5)
	$\text{DKMPP}_{\text{rbf}}$	1916.1(17.8)	1911.1(13.7)	2023.0(11.9)	2005.0(9.6)
	$\text{XGBoostPP}_{\text{p}}$	1950.2(10.3)	1938.4(6.3)	2048.1(5.1)	2025.8(4.8)
	$\text{XGBoostPP}_{\text{w}}$	1952.1(8.1)	1939.8(6.1)	2049.2(5.4)	2026.1(4.7)
			β=0.4		
10 (§4.4.2)	True	2227.0	2214.4	2128.7	2175.2
	APP_{rfm}	1955.6(39.4)	1980.7(32.3)	1916.2(30.2)	1961.8(24.3)
	$\text{DKMPP}_{\text{rbf}}$	2113.3(29.6)	2114.4(17.4)	2005.0(9.6)	2090.2(11.1)
	$\text{XGBoostPP}_{\text{p}}$	2173.2(13.0)	2163.9(12.0)	2087.4(8.1)	2128.3(8.4)
	$\text{XGBoostPP}_{\text{w}}$	2176.6(16.0)	2166.0(11.8)	2089.0(8.4)	2128.4(8.4)

the standard deviations are much smaller than those of the augmented permanental process and the deep kernel mixture point process, indicating a more stable estimation accuracy of XGBoostPP. Comparing $\text{XGBoostPP}_{\text{p}}$ with $\text{XGBoostPP}_{\text{w}}$, the latter again shows better performance on clustered processes and the improvements are, overall, slightly higher for spatially heterogeneous and clustered point patterns. However, compared to those in the lower-dimensional case, such improvements become less distinguished.

Table 4.6 Averaged integrated absolute errors (standard deviations) of different intensity estimation approaches on the Neyman-Scott process test data.

Covariates	Neyman-Scott	κ=100		κ=200	
		σ=0.02	σ=0.04	σ=0.02	σ=0.04
		β=0.5			
	KIE_{ra}	98.5(19.4)	97.8(18.1)	95.4(15.7)	94.0(14.4)
	APP_{naive}	79.6(22.7)	72.8(20.2)	69.2(17.2)	66.3(14.8)
2	APP_{rfm}	79.3(22.5)	72.9(20.0)	69.3(17.1)	66.8(14.5)
(§4.4.1)	$DKMPP_{rbf}$	148.3(21.6)	144.5(23.2)	142.5(21.8)	137.5(24.8)
	$XGBoostPP_{p}$	103.9(19.3)	99.9(19.0)	96.8(16.1)	94.1(14.6)
	$XGBoostPP_{w}$	95.3(20.1)	95.4(20.5)	92.4(18.4)	93.6(15.5)
	APP_{rfm}	223.3(16.0)	209.6(13.5)	206.8(11.3)	199.5(9.8)
10	$DKMPP_{rbf}$	173.3(19.8)	166.0(19.3)	156.8(15.4)	152.1(14.7)
(§4.4.3)	$XGBoostPP_{p}$	129.6(20.6)	124.6(19.2)	121.2(16.7)	119.9(15.3)
	$XGBoostPP_{w}$	119.5(17.5)	119.9(18.6)	114.7(17.2)	119.4(15.5)
		β=1.0			
	KIE_{ra}	175.3(17.4)	174.5(14.6)	173.6(13.3)	172.3(11.7)
	APP_{naive}	104.9(26.8)	93.6(22.7)	91.2(18.8)	85.9(16.1)
2	APP_{rfm}	104.9(26.8)	93.6(22.7)	91.2(18.8)	85.9(16.1)
(§4.4.1)	$DKMPP_{rbf}$	239.2(43.5)	234.4(46.4)	233.6(48.2)	230.9(50.6)
	$XGBoostPP_{p}$	129.7(25.4)	122.5(21.9)	120.6(19.7)	114.8(17.3)
	$XGBoostPP_{w}$	112.0(22.3)	117.5(22.7)	105.0(17.6)	113.7(18.1)
	APP_{rfm}	263.4(20.8)	242.6(16.5)	240.7(14.3)	230.0(11.0)
10	$DKMPP_{rbf}$	260.4(28.1)	249.4(30.7)	244.4(28.8)	235.4(34.1)
(§4.4.3)	$XGBoostPP_{p}$	172.2(22.3)	155.9(20.1)	153.1(17.1)	146.1(15.8)
	$XGBoostPP_{w}$	148.7(22.2)	153.7(20.1)	138.8(17.9)	146.1(15.8)
		β=0.2			
	APP_{rfm}	221.1(13.8)	211.0(11.6)	209.0(10.6)	207.3(11.4)
10	$DKMPP_{rbf}$	164.7(22.2)	155.4(19.4)	146.4(15.5)	141.6(14.4)
(§4.4.2)	$XGBoostPP_{p}$	120.0(19.4)	114.8(16.2)	110.0(13.3)	110.9(13.4)
	$XGBoostPP_{w}$	113.9(17.0)	111.2(15.3)	104.9(13.1)	109.9(13.1)
		β=0.4			
	APP_{rfm}	253.9(14.7)	241.1(12.0)	238.7(11.1)	232.7(9.2)
10	$DKMPP_{rbf}$	212.9(18.2)	204.3(18.6)	196.8(16.6)	191.9(15.4)
(§4.4.2)	$XGBoostPP_{p}$	161.3(17.7)	158.1(16.8)	143.5(13.4)	153.3(14.2)
	$XGBoostPP_{w}$	143.6(17.5)	153.0(17.8)	138.4(12.9)	153.2(14.2)

4.4.3 Simple Intensity with Many Nuisance Variables

As an extensive experiment, we simulate data of the same models with the simple intensity function $\lambda(\boldsymbol{s}) = \alpha \exp\{\beta[z_1(\boldsymbol{s}) + z_2(\boldsymbol{s})]\}$ as in Section 4.4.1, however, input the ten covariates from Section 4.4.2, $z_1(\boldsymbol{s}), \ldots, z_{10}(\boldsymbol{s})$, for intensity estimation, thus with a large number of nuisance variables. Such a scenario occurs widely in practice when numerous covariates are available while researchers are uncertain about which ones are most relevant and should be included, e.g. in the situation of chimney fire prediction in Chapter 2.

We again evaluate all approaches with the two performance metrics and report the results. From Tables 4.1– 4.6, it is obvious that XGBoostPP exhibits

more robust behavior compared to the augmented permanental process and the deep kernel mixture point process, which can be evidenced by the significantly smaller reductions in the Poisson log-likelihoods and smaller increases in the integrated absolute errors when comparing the results to those of Section 4.4.2. Such robustness suggests great potentials of XGBoostPP in practical applications of classic medium-sized point pattern data. In addition, the reductions and increases of the deep kernel mixture point process is less than the augmented permanental process, revealing that, while the true intensity function is simple, the augmented permanental process is less robust when there is considerable uncertainty in selecting the most relevant covariates.

4.4.4 Poisson Toy Examples

In addition to the main simulation study, we test XGBoostPP on three Poisson toy examples to intuitively visualize its flexibility in modelling various nonlinear relationships between covariates and the intensity function. The examples are also created on $\mathcal{W} = [0,1]^2$ but take the spatial x, y coordinates as covariates. The first example has the intensity function $\lambda(x,y) = \exp(3 + 2x + 7y)$, which is simply log-linear. The second example employs the intensity $\lambda(x,y) = \exp[8 + 16(x - 0.5)^2]$, which aims to show the automatic variable selection capability of XGBoostPP. The third example considers the intensity function $\lambda(x,y) = \exp[2 + 4\sin(16x) + 6\sin(16y)]$, which intends to show the capacity for modelling extraordinary covariate responses.

For all examples, we generate a point pattern based on the designed intensity function and fit XGBoostPP using the Poisson likelihood loss function to estimate the intensity as they are all Poisson processes. To test the flexibility, we only input x, y to XGBoostPP and force it to learn the covariate relationships and the intensity trends on its own. We plot point patterns, true intensities and the estimated intensities by XGBoostPP in Figure 4.1. The plots clearly show that our XGBoostPP model can detect different covariate relationships and perform an automatic variable selection.

4.5 Real Data Analyses

To demonstrate the utility of XGBoostPP in practice, we apply it to two real data sets: tropical forest data on Barro Colorado Island in Panama and kitchen fire data in the Twente region in the Netherlands.

For each data set, we conduct a four-fold cross-validation to evaluate different intensity estimation approaches based on the test Poisson log-likelihood (i.e., we randomly split the data points into four subsets, assign one for test and the others for training). Based on the four-fold cross-validation, the test Poisson log-likelihood over the data of one of the four subsets $\mathbf{x}_i$, with $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_4\}$,

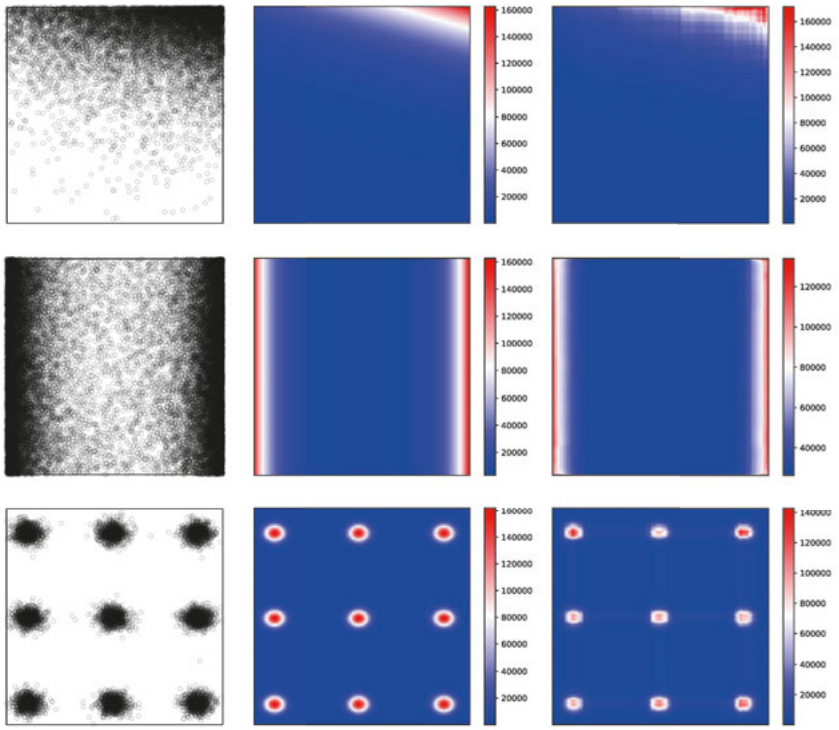


Figure 4.1 Point patterns, true intensities and estimated intensities by XGboostPP$_{\text{p}}$ for the three Poisson toy examples.

reads

$$PL_{\text{test},i} = \sum_{\boldsymbol{x} \in \mathbf{x}_i} \log \left[\frac{1}{3} \widehat{\lambda}_i(\boldsymbol{x}) \right] - \frac{1}{3} \int_{\mathcal{W}} \widehat{\lambda}_i(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}, \tag{4.25}$$

where $\widehat{\lambda}_i(\boldsymbol{s})$ is the estimated intensity function for fold i using the other three folds of training data. We sum up this test Poisson log-likelihood over all four subsets of $\mathbf{x}_i$ as the cross-validated test Poisson log-likelihood

$$PL_{\text{test}} = \sum_{i=1}^{4} PL_{\text{test},i} = \sum_{i=1}^{4} \left\{ \sum_{\boldsymbol{x} \in \mathbf{x}_i} \log \left[\frac{1}{3} \widehat{\lambda}_i(\boldsymbol{x}) \right] - \frac{1}{3} \int_{\mathcal{W}} \widehat{\lambda}_i(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} \right\}. \tag{4.26}$$

4.5.1 Tropical Forestry Data

The tropical forest data set documents the spatial distributions of various tree species across the 50-hectare Barro Colorado Island over the past 30 years (Condit et al., 2019). We select two tree species to study here: the Beilschmiedia pendula (Bei) and the Capparis frondosa (Capp), containing 3604 and 3299 tree locations, respectively. We investigate eight covariates that may explain their spatial distributions, including terrain elevation and slope, four soil nutrients and solar and wetness indices. Details of these covariates are listed in Table 4.7.

Table 4.7 The covariates involved in the tropical forestry data set.

Data	Covariate	Description
	Elev	The terrain elevation
	Grad	The terrain slope
	Cu	The content of Cu
Bei	Nmin	The content of Nmin
Capp	P	The content of P
	pH	The pH value of soil
	Solar	The solar index
	Twi	The wetness index

We report the cross-validated test Poisson log-likelihoods in Table 4.9. It indicates that, on both the Beilschmiedia pendula and the Capparis frondosa data, XGBoostPP outperforms the augmented permanental process and the deep kernel mixture point process. Since the Beilschmiedia pendula species appears more clustered compared to the Capparis frondosa (Guan & Shen, 2010; Yue & Loh, 2011), XGBoostPP$_{\mathrm{w}}$ improves the performance on the former relatively more. Figure 4.2 displays the estimated log-intensities by the augmented permanental process and XGBoostPP, showing that, in comparison to the former, XGBoostPP produces significantly different estimates for the areas with a small number of point observations.

4.5.2 Kitchen Fire Data

The kitchen fire data (Fire) comprises 699 kitchen fire incidents in Twente from 2004 to 2020 (cf., Section 2.7.1). The covariates of interest include building information, urbanity degrees, population components and energy consumption, leading to 29 covariates in total. Details of these covariates are listed in Table 4.8.

Table 4.9 provides the cross-validated test Poisson log-likelihoods results for XGBoostPP only, as the other two approaches encounter computational problems and produce significantly less accurate estimates. We also plot the fire data and the estimated log-intensities by XGBoostPP in Figure 4.3. On this data set, XGBoostPP demonstrates its strong capacity to handle covariate spaces of high dimensions given very limited training data. As expected, XGBoostPP$_{\mathrm{w}}$ again yields a higher test Poisson log-likelihood than XGBoostPP$_{\mathrm{p}}$ due to the apparent spatial dependence observed in the raw data.

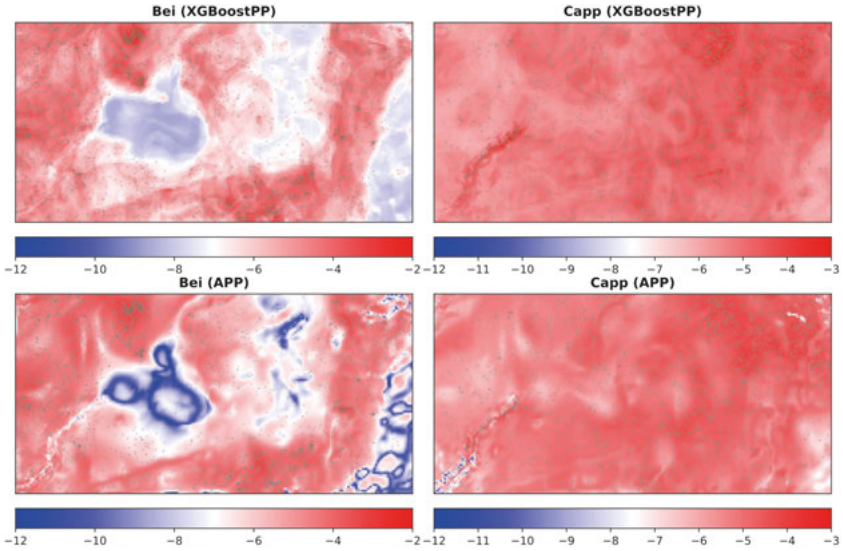


Figure 4.2 Observed tree locations of Beilschmiedia pendula and Capparis frondosa and estimated log-intensities by XGBoostPP$_w$ and the augmented permanental process (APP).

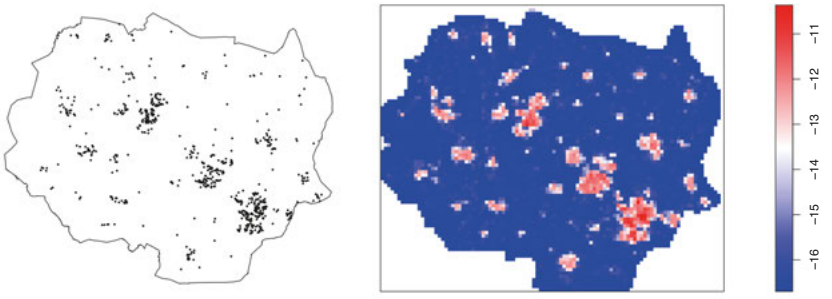


Figure 4.3 Kitchen fire locations and estimated log-intensities by XGBoostPP$_w$.

4.5.3 Practical Utility

The estimated intensities by XGBoostPP better reflect the distribution of point occurrences than the stat-of-the-art approaches, which can help forestry scientists learn the suitable living environment for tree species and can provide firefighters with the risk information for proper personnel and equipment arrangements. Moreover, XGBoostPP can accommodate more available covariates, providing prior knowledge for further analyzing the relationship between a putative variable and the intensity function. For instance, forestry researcher may thus be

Table 4.8 The covariates involved in the kitchen fire data set.

Data	Covariate	Description
Fire	House	The total number of houses
	House_indu	The number of houses with an industrial function
	House_hotl	The number of houses with a hotel function
	House_resi	The number of houses with a residential function
	House_20	The number of houses constructed before 1920
	House_2045	The number of houses constructed in $[1920, 1945)$
	House_4570	The number of houses constructed in $[1945, 1970)$
	House_7080	The number of houses constructed in $[1970, 1980)$
	House_8090	The number of houses constructed in $[1980, 1990)$
	House_90	The number of houses constructed after 1990
	House_frsd	The number of free standing houses
	House_other	The number of other houses
	Resid	The number of residents
	Resid_14	The number of residents with an age in $[0, 14)$
	Resid_1524	The number of residents with an age in $[15, 24)$
	Resid_2544	The number of residents with an age in $[25, 44)$
	Resid_4564	The number of residents with an age in $[45, 64)$
	Resid_65	The number of residents with an age over 65
	Man	The number of male residents
	Woman	The number of female residents
	Resid_frsd	The number of residents living in free standing houses
	Address	The density of addresses in the block
	Urbanity	The urbanity of the block
	Town	Boolean variable indicating the presence of a town
	Poor	The percentage of poor residents (income $0-20$ percent)
	Rich	The percentage of rich residents (income $80-100$ percent)
	Value_house	The average value of the houses in the block
	Gas_use	The average gas use in m^3 in the block
	Elec_use	The average electricity use in kWh in the block

Table 4.9 Cross-validated (4-fold) test Poisson log-likelihoods of different intensity estimation approaches on tropical forestry and kitchen fire data.

Data set	Bei	Capp	Fire
Covariates	8	8	29
APP_{rfm}	-24955.4	-24295.8	-
$\text{DKMPP}_{\text{rbf}}$	-26505.6	-24487.8	-
$\text{XGBoostPP}_{\text{p}}$	-24557.6	-24097.3	-10773.2
$\text{XGBoostPP}_{\text{w}}$	-24376.1	-24080.5	-10519.7

motivated to investigate the specific influence of a nutrient on the growth of a tree species, and firefighters may use the information to organize public campaigns for fire prevention.

4.6 Summary

In this chapter, we proposed a novel tree-based ensemble method, XGBoostPP, to nonparametrically estimate the intensity of a point process as a function of covariates, working extremely well for classic medium-sized data. Two loss functions

were carefully designed for model estimation, the second of which was especially derived for our tree-based ensemble model to improve the estimation efficiency for clustered processes. We also developed an efficient learning algorithm for model fitting and proposed an associated validation procedure to select hyperparameters. Numerical and real data analyses show that XGBoostPP achieves superior performance to state-of-the-art approaches, suggesting it a useful and flexible tool for analyzing unknown, complex medium-sized point patterns in practice.

However, one limitation of the proposed approach is the computational demand of hyperparameter selection due to the combination of the greedy search algorithm and the cross-validation procedure (if one has many hyperparameter candidates). In experiments, by using parallel computing tools to tune hyperparameters, XGBoostPP has a slightly larger running time than the augmented point process and a smaller running time than the deep kernel mixture point process.

For future work, the first direction would be to improve the estimation of Ripley's K-function for point patterns with high-dimensional covariate spaces, so that the approximated dynamic weighted likelihood loss could be improved. Moreover, it would also be interesting to extend the proposed method to estimating second-order intensities for non-stationary point pattern data.

CHAPTER 5

Bayesian Inference for Independent Cluster Point Processes

5.1 Introduction

Neyman-Scott point processes (Neyman & Scott, 1958) have been widely used to model spatial and spatio-temporal clustered point patterns. A Neyman-Scott process consists of two layers: the first layer is a Poisson point process that introduces latent cluster centres, often referred to as parents; secondly, around cluster centres, independent Poisson processes are generated to produce offspring points per cluster. Due to the straightforward probability density function of Poisson processes, statistical inference for Neyman-Scott processes has been well established, encompassing both frequentist and Bayesian approaches. Frequentist methods primarily focus on moment-based estimation (Guan, 2006; Waagepetersen, 2007; Tanaka, Ogata & Stoyan, 2008; Waagepetersen & Guan, 2009; Mrkvička, Muška & Kubečka, 2012; Prokešová & Vedel Jensen, 2013; Tanaka & Ogata, 2014; Baddeley, Rubak & Turner, 2015; Baddeley et al., 2022). For instance, first-order estimating equations are used to estimate the coefficients related to spatial and temporal heterogeneity, and minimum contrast estimation is performed based on second-order statistics, such as Ripley's K-function or the pair correlation function (cf., Section 1.1), to optimize the shape parameters of a cluster (i.e., describing the spread of offspring points in it). On the Bayesian side, inference is more challenging since neither the number of latent cluster centres nor their locations are known. To address this trans-dimensional problem, reversible jump Markov chain Monte Carlo algorithms (MCMC) (Green, 1995, 2003), equipped with birth-and-death moves, have been developed (Van Lieshout & Baddeley, 2002; Møller & Waagepetersen, 2004, 2007; Mrkvička, Muška & Kubečka, 2012). Compared to frequentist methods, Bayesian methods not only enable parameter estimation but also allow for a more comprehensive inference on the overall doubly stochastic model.

A well-known limitation of Neyman-Scott processes in modelling clustered point patterns is the assumption of a Poisson point process for latent cluster centres, which may cause an overestimation of clusters when fitted to real data. The problem arises since treating each offspring point as a single-event cluster

could result in a higher posterior probability of the model. To address this, a generalization has been proposed, replacing the Poisson prior with a repulsive point process, leading to a family of Cox cluster point processes (Daley & Vere-Jones, 2009). A typical example is the use of a Strauss process (Strauss, 1975) to introduce regularity among cluster centres. In this context, Bayesian inference approaches are particularly suitable due to the Markovian properties of Strauss processes: Van Lieshout & Baddeley (2002) employed a spatial birth-and-death process to sample latent cluster centres and applied the Monte Carlo maximum likelihood method to estimate the intensity and shape parameters of clusters, while the parameters in the Strauss prior were treated as fixed in practice. For frequentist approaches, some general results for summary statistics have been derived (Møller & Torrisi, 2005), however, their formulas involve intractable high-dimensional integrals. To overcome the computational challenges posed by the Strauss prior, Yau & Loh (2012) developed a closed-form approximation of Ripley's K-function, which again allows the model to be fitted using minimum contrast estimation.

Although the Poisson prior has been generalized to a broader class of point processes, Cox cluster process models are still limited in many applications. Two key issues persist: (i) the Poisson assumption on the structure of a cluster remains limited and may not adequately capture the complexity of real data; (ii) when the evidence of clustering in data observations is weak, classic model fitting procedures can become unstable and unreliable, as pointed out by Baddeley et al. (2022). Such challenges are prevalent in practical scenarios, for instance, in a risk prediction study of car arson fires that motivates our work. According to firefighters, car arson fires are driven by activities of specific groups of arsonists and tend to occur in clusters. A cluster can be identified when there is at least one fire incident reported. Arson attempts may also occur but are not registered, as the interest of the fire service is to detect emerging arson clusters and implement immediate measures to put out actual fires. Due to inherent positivity, the number of offspring points in a cluster does not follow a Poisson distribution as in Cox cluster processes. Additionally, joint crime records from the police indicate that only about 10% of arsonists are repeat offenders, resulting in a large proportion of single-event arson clusters (cf., Figure 5.1(right)). Consequently, the clustering strength is rather weak and is difficult to learn. To effectively model and fit this kind of data, a more general framework – independent cluster point processes (Daley & Vere-Jones, 2009), where both parent and offspring processes are not restricted to the Poisson assumption – is required, and an inference method to flexibly and stably capture data complexity is essential.

The body of literature addressing inference for independent cluster point processes is quite small. Van Lieshout (1994) designed deterministic and stochastic algorithms to sample the cluster information from complicated posterior distributions. Swain & Clark (2011) proposed a Bayesian method to estimate the intensity function of a special class of independent cluster models whose parent processes contain only one cluster. The most recent related work is that con-

cerned with Bayesian mixture models, which extend traditional finite mixture models (Everitt & Hand, 1981) by assuming explicit prior distributions on the number of clusters (Richardson & Green, 1997; Miller & Harrison, 2018). To reduce redundant cluster components, repulsive priors, as those used in Cox cluster processes (Van Lieshout & Baddeley, 2002; Yau & Loh, 2012), were incorporated to enforce regularity among cluster centres, and MCMC methods, such as Gibbs sampling algorithms, were developed for efficient model inference (Xie & Xu, 2020; Beraha et al., 2022). Although Bayesian mixture models share similarities with independent cluster processes, the latter class is better-suited to modelling spatial or spatio-temporal data for two reasons: (i) the number of clusters in an independent cluster process is governed by the intensity function of the parent process, which can be tuned easily to the spatial or spatio-temporal domain of interest; (ii) point process models have a well-defined probabilistic structure, treating both cluster centres and offspring points as random variables, whereas mixture models estimate clusters conditional on fixed point observations. Notably, Wang et al. (2023) built a theoretical connection between Neyman-Scott processes and Bayesian mixture models and adapted the standard collapsed Gibbs sampling algorithm to infer the former with Gamma weights.

In this study, we continue to explore the independent cluster point process as a framework for modelling spatial and spatio-temporal data and develop a two-step Bayesian approach for its statistical inference. The proposed method offers significant flexibility for introducing interactions between cluster centres and accommodating customized cluster structures, provided that their probability functions have an analytical form. As a specific example, we present a structure for modelling the arson fire data mentioned above, where the prior is set to a Strauss process and the number of offspring points in each cluster follows a shifted-Poisson distribution. We derive the posterior probability components for this model structure and conduct numerical experiments on it across various parameter settings to validate our Bayesian inference approach. In a follow-up step, we will demonstrate the practical utility of the proposed method by applying it to the real fire data and highlighting its effectiveness in predicting emerging clusters in a spatio-temporal manner.

The remainder of this chapter is organized as follows. Section 5.2 describes the framework of independent cluster point processes in detail. Section 5.3 introduces our proposed Bayesian inference approach. In Section 5.4, we present a novel model structure for arson fire data, followed by a simulation study in Section 5.5 to validate the inference method. Section 5.6 discusses the application of our approach to real fire data to demonstrate its applicability. Finally, the chapter concludes with a summary.

5.2 The Independent Cluster Point Process

An independent cluster point process can be mathematically formulated in the language of a marked point process (Daley & Vere-Jones, 2009). The ground

process of points, representing cluster centres, is typically modelled by a Markov point process (Ripley & Kelly, 1977; Van Lieshout, 2000), leveraging its analytical probability density with respect to a unit-rate Poisson process on the same observation window. The mark kernel generates events for each cluster through first using a probability mass function to determine the number of offspring points, followed by a scatter allocation function to assign their locations.

In this section, we start with revisiting the structures of the ground process and the mark kernel. Then, we integrate these components to obtain the overall independent cluster process model and introduce its probability framework.

5.2.1 The Ground Process

In principle, the ground process can be any locally stable Markov point process. Let us denote the ground process by X, defined on some space-time window $\mathcal{S}^+ \subset \mathbb{R}^2 \times \mathbb{R}$, with a probability density $P_X(\cdot)$ that is absolutely continuous with respect to a unit-rate Poisson point process on $\mathcal{S}^+$. Let $\mathbf{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ represent a realization of X, thus $\mathbf{x} \in \exp[\mathcal{S}^+]$. The process X is said to be a Markov point process at range (r_s, r_t) if the ratio

$$\lambda_X(\boldsymbol{\xi}|\mathbf{x}) = \frac{P_X(\mathbf{x} \cup \{\boldsymbol{\xi}\})}{P_X(\mathbf{x})} \tag{5.1}$$

is well-defined for all $\boldsymbol{\xi} \in \mathcal{S}^+$ and depends only on those $\boldsymbol{x} \in \mathbf{x}$ for which $\|\boldsymbol{x}_s - \boldsymbol{x}_s\|_2 < r_s$ and $|\boldsymbol{x}_t - \boldsymbol{x}_t| < r_t$, where $\boldsymbol{x}_s$ and $\boldsymbol{x}_t$ represent the spatial and temporal coordinates of $\boldsymbol{x}$, respectively, and r_s, r_t are related interaction distances. Moreover, this ratio, also known as the conditional intensity function of X, must be uniformly bounded in both of its arguments to ensure the local stability of the ground process.

As mentioned in Section 5.1, to avoid an overestimation of the number of clusters in analyzing real data, it is useful to assume X to be a repulsive point process. A common example is the pairwise interaction process (Ripley & Kelly, 1977), which is characterized by some trend function $b_{\boldsymbol{\alpha}}(\cdot) : \mathcal{S}^+ \mapsto \mathbb{R}^+$ and interaction function $h_{\boldsymbol{\alpha}}(\cdot,\cdot) : \mathcal{S}^+ \times \mathcal{S}^+ \mapsto \mathbb{R}^+$, parameterized by $\boldsymbol{\alpha}$. This process has the following probability density with respect to a unit-rate Poisson process on $\mathcal{S}^+$

$$P_X(\mathbf{x}) = \frac{1}{C(\boldsymbol{\alpha})} \prod_{i=1}^{m} b_{\boldsymbol{\alpha}}(\boldsymbol{x}_i) \prod_{1 \le j < k \le m} h_{\boldsymbol{\alpha}}(\boldsymbol{x}_j, \boldsymbol{x}_k), \tag{5.2}$$

where $C(\boldsymbol{\alpha})$ is a normalizing constant and m denotes the number of points in $\mathbf{x}$. Intuitively, the trend function $b_{\boldsymbol{\alpha}}(\cdot)$, if not constant, modulates the intensity of points in X by capturing the spatial and temporal heterogeneity of cluster centres, while the interaction function $h_{\boldsymbol{\alpha}}(\cdot)$ introduces interactions between these centres.

5.2.2 The Mark Kernel

For an independent cluster process, the mark kernel is independently applied to every cluster centre $\boldsymbol{x} \in X$ to produce corresponding offspring points. Denote the mark of the cluster centred at $\boldsymbol{x}_i$ by Y_i, defined on a slightly smaller window $\mathcal{S} \subset \mathcal{S}^+ \subset \mathbb{R}^2 \times \mathbb{R}$. This setup allows a cluster centred outside $\mathcal{S}$ but within $\mathcal{S}^+$ to have offspring points falling inside $\mathcal{S}$. Let $\mathbf{y}_i = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_i}\}$ represent a realization of Y_i, thus $\mathbf{y}_{\boldsymbol{i}} \in \exp[\mathcal{S}]$. The probability density of Y_i with respect to a unit-rate Poisson process on $\mathcal{S}$ is given by

$$g(\mathbf{y}_i|\boldsymbol{x}_i) = \exp(|\mathcal{S}|)P(n_i|\boldsymbol{x}_i)n_i!\frac{\prod_{\boldsymbol{u}\in\mathbf{y}_i} k(\boldsymbol{u}|\boldsymbol{x}_i)}{\left[\int_{\mathcal{S}} k(\boldsymbol{s}|\boldsymbol{x}_i)\mathrm{d}\boldsymbol{s}\right]^{n_i}}. \tag{5.3}$$

Here, n_i is the number of offspring points, $P(n_i|\boldsymbol{x}_i)$ is the probability of observing n_i points in Y_i, $k(\cdot|\boldsymbol{x}_i)$ is a scatter allocation kernel centred at $\boldsymbol{x}_i$ that describes the spatial and temporal spread of offspring points, and $|\mathcal{S}|$ is the volume of $\mathcal{S}$.

In general, the probability mass function $P(\cdot|\boldsymbol{x}_i)$ and the scatter allocation kernel $k(\cdot|\boldsymbol{x}_i)$ can be chosen freely, provided that (5.3) is integrable. For instance, consider a cluster Y_i centred at $\boldsymbol{x}_i$ with a total of a_i offspring points in $\mathbb{R}^2 \times \mathbb{R}$, where a_i follows a probability mass function $f(a) \geq 0$ and $\sum_a f(a) = 1$. The probability of observing n_i offspring points within $\mathcal{S} \subset \mathbb{R}^2 \times \mathbb{R}$ is given by the condition that the cluster contains at least n_i offspring points, with exactly n_i of them falling in $\mathcal{S}$ and the remaining $a_i - n_i$ in the complement. Since only offspring points within $\mathcal{S}$ are observable, summing over all possible values of a_i leads to

$$P(n_i|\boldsymbol{x}_i) = \sum_{a_i=n_i}^{\infty} f(a_i)\binom{a_i}{n_i} q(\boldsymbol{x}_i)^{n_i}\left[1 - q(\boldsymbol{x}_i)\right]^{a_i-n_i}, \tag{5.4}$$

where $q(\boldsymbol{x}_i)$ is the success probability of the binomial distribution. Following (5.3), we have $q(\boldsymbol{x}_i) = \int_{\mathcal{S}} k(\boldsymbol{s}|\boldsymbol{x}_i)\mathrm{d}\boldsymbol{s}$. The symmetric joint conditional probability of assigning the n_i points to specific locations of $\mathbf{y}_i$ is thus given by $\prod_{\boldsymbol{s}\in\mathbf{y}_i}[k(\boldsymbol{s}|\boldsymbol{x}_i)/q(\boldsymbol{x}_i)]$. Then, (5.3) becomes

$$g(\mathbf{y}_i|\boldsymbol{x}_i) = \exp(|\mathcal{S}|)\prod_{\boldsymbol{s}\in\mathbf{y}_i} k(\boldsymbol{s}|\boldsymbol{x}_i)\sum_{a_i=n_i}^{\infty} f(a_i)\frac{a_i!}{(a_i-n_i)!}\left[1 - q(\boldsymbol{x}_i)\right]^{a_i-n_i}. \tag{5.5}$$

Setting $c_i = a_i - n_i \geq 0$, we can rewrite (5.5) as

$$g(\mathbf{y}_i|\boldsymbol{x}_i) = \exp(|\mathcal{S}|)\prod_{\boldsymbol{s}\in\mathbf{y}_i} k(\boldsymbol{s}|\boldsymbol{x}_i)\sum_{c_i=0}^{\infty} f(c_i+n_i)\frac{(c_i+n_i)!}{c_i!}\left[1 - q(\boldsymbol{x}_i)\right]^{c_i}. \tag{5.6}$$

Cox cluster processes assume $f(\cdot)$ to be a Poisson mass function with some mean $\mu > 0$, so that $f(a) = \mu^a e^{-\mu}/a!$ for $a \in \mathbb{N}$. Substituting this into (5.6) and

recalling the power series of the exponential function, we obtain

$$\begin{aligned} g(\mathbf{y}_i|\boldsymbol{x}_i) &= \exp(|\mathcal{S}| - \mu)\mu^{n_i} \prod_{\boldsymbol{s}\in\mathbf{y}_i} k(\boldsymbol{s}|\boldsymbol{x}_i) \sum_{c_i=0}^{\infty} \frac{\{\mu\,[1-q(\boldsymbol{x}_i)]\}^{c_i}}{c_i!} \\ &= \exp[|S| - \mu q(\boldsymbol{x}_i)] \prod_{\boldsymbol{s}\in\mathbf{y}_i} \mu k(\boldsymbol{s}|\boldsymbol{x}_i), \end{aligned} \tag{5.7}$$

which is indeed the probability density for Y_i to be a Poisson point process on $\mathcal{S}$ with intensity function $\mu k(\boldsymbol{s}|\boldsymbol{x}_i)$. Classic choices for $k(\cdot|\boldsymbol{x}_i)$ include the disc and the Gaussian kernels. The disc kernel is defined as $k(\boldsymbol{s}|\boldsymbol{x}_i) = 1$ if the distance between $\boldsymbol{s}$ and $\boldsymbol{x}_i$ is within a specified range, and $k(\boldsymbol{s}|\boldsymbol{x}_i) = 0$ otherwise. The Gaussian kernel reads

$$k(\boldsymbol{s}|\boldsymbol{x}_i) = \frac{1}{(2\pi)^{3/2}\det(\boldsymbol{\Sigma})} \exp\left[-\frac{1}{2}(\boldsymbol{s}-\boldsymbol{x}_i)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{x}_i)\right], \tag{5.8}$$

where $\boldsymbol{\Sigma}$ is the spatio-temporal covariance matrix and $\det(\boldsymbol{\Sigma})$ is its determinant.

Such a combination of the probability mass function $P(\cdot|\boldsymbol{x}_i)$ and the scatter allocation kernel $k(\cdot|\boldsymbol{x}_i)$ under the structure of (5.3) defines a class of finite point processes that assume independence among points. Indeed, this structure can be extended to point processes involving point interactions by replacing the terms in (5.3) with products of conditional intensity functions, e.g. (5.1) for a Markov point process. More general models allow the mark kernel to be specific to its cluster centre $\boldsymbol{x}_i$, for instance, changing the mean parameter μ in (5.7) to μ_i that depends on $\boldsymbol{x}_i$. However, it is worth noting that our proposed Bayesian inference framework will remain applicable despite these modifications to the explicit form of (5.3), ensuring flexibility in handling general independent cluster process models.

5.2.3 The Independent Cluster Model

Combining the ground process and the mark kernel defined in Sections 5.2.1 and 5.2.3, we obtain the complete independent cluster model.

Let a realization of this complete process be denoted by $\{(\boldsymbol{x}_i, \mathbf{y}_i)\}$, with $i = 1, \ldots, n\{\mathbf{x}\}$, where $n\{\cdot\}$ denotes the number of points in a set. Each $(\boldsymbol{x}_i, \mathbf{y}_i)$ then resides in $\mathcal{S}^+ \times \exp[\mathcal{S}]$. Since the processes Y_i are assumed to be conditionally independent given $\mathbf{x}$, we can directly aggregate all $\mathbf{y}_i$'s to collect the overall process of offspring points. With respect to the $n\{\mathbf{x}\}$-fold product of unit-rate Poisson processes defined on $\mathcal{S}$, the aggregated process has the following conditional probability

$$P_{Y_1,\ldots,Y_{n\{\mathbf{x}\}}|X}(\mathbf{y}_1, \cdots, \mathbf{y}_{n\{\mathbf{x}\}}|\mathbf{x}) = \exp(|\mathcal{S}|) \prod_{i=1}^{n\{\mathbf{x}\}} \left[g(\mathbf{y}_i|\boldsymbol{x}_i)\exp(-|\mathcal{S}|)\right]. \tag{5.9}$$

In practice, only the offspring points are observed, without any information regarding the underlying cluster assignments. Let U represent the process of

all offspring points, with a realization denoted by $\mathbf{u} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$, so that $\mathbf{u} \in \exp[\mathcal{S}]$. It is important to note that different partitions of $\mathbf{y}_1, \cdots, \mathbf{y}_{n\{\mathbf{x}\}}$ can result in the same realization of $\mathbf{u}$. For ease of notation, we consider a mapping to assign an offspring point to its parent cluster and denote it by $\psi_{\mathbf{x}} : \{1, \ldots, n\} \to \{1, \ldots, n\{\mathbf{x}\}\}$. It uniquely specifies the marks $\mathbf{y}_1, \cdots, \mathbf{y}_{n\{\mathbf{x}\}}$ over $\mathbf{x}$ given $\mathbf{u}$. In the remainder of this chapter, we use $\psi_{\mathbf{x}}^{-1}(\cdot)$ to denote the inverse of $\psi_{\mathbf{x}}(\cdot)$ and $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(i)}\}$ to collect the offspring points of the cluster centred at $\boldsymbol{x}_i$ from $\mathbf{u}$ by setting the constraint: $\mathbf{u} = \{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(1)}\} \cup \cdots \cup \{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(n\{\mathbf{x}\})}\}$. Corresponding to (5.9), the conditional probability of observing $\mathbf{u}$ given $\mathbf{x}$ reads

$$P_{U|X}(\mathbf{u}|\mathbf{x}) = \exp(|\mathcal{S}|) \sum_{\psi_{\mathbf{x}}} \prod_{i=1}^{n\{\mathbf{x}\}} \left[g\left(\left\{ \boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(i)} \right\} | \boldsymbol{x}_i \right) \exp(-|\mathcal{S}|) \right], \tag{5.10}$$

again with respect to the $n\{\mathbf{x}\}$-fold product of unit-rate Poisson processes defined on $\mathcal{S}$.

When the processes Y_i are Poisson, the independent cluster process reduces to a Cox cluster process, yielding that

$$\begin{aligned} P_{U|X}(\mathbf{u}|\mathbf{x}) &= \exp(|\mathcal{S}|) \sum_{\psi_{\mathbf{x}}} \prod_{i=1}^{n\{\mathbf{x}\}} \left\{ \exp\left[-\mu q(\boldsymbol{x}_i)\right] \prod_{\boldsymbol{s} \in \{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(i)}\}} \mu k(\boldsymbol{s}|\boldsymbol{x}_i) \right\} \\ &= \exp(|\mathcal{S}|) \exp\left[-\sum_{i=1}^{n\{\mathbf{x}\}} \mu q(\boldsymbol{x}_i) \right] \prod_{\boldsymbol{s} \in \mathbf{u}} \left[\sum_{i=1}^{n\{\mathbf{x}\}} \mu k(\boldsymbol{s}|\boldsymbol{x}_i) \right]. \end{aligned} \tag{5.11}$$

This simplification removes the dependence on $\psi_{\mathbf{x}}$, implying that for Cox cluster processes, U, conditional on X, is a superposition of independent Poisson processes Y_i, which is again a Poisson process with the summed intensity function $\sum_{i=1}^{n\{\mathbf{x}\}} \mu k(\boldsymbol{s}|\boldsymbol{x}_i)$. Various frequestist (e.g., Waagepetersen & Guan, 2009; Yau & Loh, 2012) and Bayesian (e.g., Van Lieshout & Baddeley, 2002; Møller & Waagepetersen, 2004, 2007) approaches have been developed to infer such models based on their elegant, analytical moment and probability expressions.

When the processes Y_i are non-Poisson, $g(\cdot|\boldsymbol{x}_i)$ may include additional terms that depend explicitly on the partition mapping $\psi_{\mathbf{x}}$ (see, e.g. the model structure designed for arson fire data in Section 5.4), preventing the simplification achieved in (5.11). In these cases, standard statistical inference approaches, such as estimating equations, composite likelihood estimation and minimum contrast estimation (Guan, 2006; Waagepetersen, 2007; Mrkvička, Muška & Kubečka, 2012), become difficult or intractable when deriving the closed-form expressions of first- and second-order properties.

To enable flexible inference for a general independent cluster point process given only the overall offspring observations $\mathbf{u}$, we decide to build a Bayesian estimation framework by focusing on the following two terms determining the

posterior distribution of the overall model:

$$P(\mathbf{x}|\mathbf{u}) = \frac{1}{C(\mathbf{u})} P(\mathbf{x}) P(\mathbf{u}|\mathbf{x}), \tag{5.12}$$

and

$$P(\psi_{\mathbf{x}}|\mathbf{x}, \mathbf{u}) = P(\psi_{\mathbf{x}}, \mathbf{u}|\mathbf{x}) / P(\mathbf{u}|\mathbf{x}), \tag{5.13}$$

where $C(\mathbf{u})$ is a normalizing constant, and recalling (5.10)

$$P(\psi_{\mathbf{x}}, \mathbf{u}|\mathbf{x}) = \exp(|\mathcal{S}|) \prod_{i=1}^{n\{\mathbf{x}\}} \left[g\left(\left\{ \boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(i)} \right\} | \boldsymbol{x}_i \right) \exp(-|\mathcal{S}|) \right]. \tag{5.14}$$

In the case of a pairwise interaction point process defined by (5.2) as the ground process,

$$P(\mathbf{x}|\mathbf{u}) = \frac{1}{C(\boldsymbol{\alpha}|\mathbf{u})} \prod_{i=1}^{n\{\mathbf{x}\}} b_{\boldsymbol{\alpha}}(\boldsymbol{x}_i) \prod_{1 \le j < k \le n\{\mathbf{x}\}} h_{\boldsymbol{\alpha}}(\boldsymbol{x}_j, \boldsymbol{x}_k) P(\mathbf{u}|\mathbf{x}). \tag{5.15}$$

Note that the normalizing constant in (5.2) now becomes $C(\boldsymbol{\alpha}|\mathbf{u})$.

For a complete inference, we will need to estimate both state information, including the locations of cluster centres $\mathbf{x}$ and the mapping function $\psi_{\mathbf{x}}(\cdot)$, as well as model parameters, including the trend and interaction parameters $\boldsymbol{\alpha}$ in the ground process and the intensity and shape parameters $\mu, \boldsymbol{\Sigma}$ of a cluster, simultaneously. Moreover, it is worth pointing out that the number of latent cluster centres $n\{\mathbf{x}\}$ varies, yielding the trans-dimensional challenge encountered when sampling from doubly stochastic point processes.

5.3 A Two-step Bayesian Inference Approach

In this section, we develop a two-step Bayesian approach to estimate the independent cluster point process model introduced in Section 5.2 based on (5.12) and (5.13). Specifically, in the first step, we propose a Markov chain Monte Carlo (MCMC) method to sample the state information of the model given certain parameters. In the second step, we develop a Monte Carlo expectation-maximization (MCEM) algorithm to optimize model parameters iteratively, with the sampled states.

5.3.1 State Estimation by Markov chain Monte Carlo

Reversible jump MCMC algorithms (Green, 1995, 2003) have been used to address the trans-dimensional problem when sampling from doubly stochastic point processes (Møller & Waagepetersen, 2004). For the independent cluster model, similar ideas can be considered, however, adjustments are necessary due to more complex formulas related to the posterior probability. Below, we first describe

the design of our Markov chain and then introduce the specific moves we propose for efficiently sampling the states from the posterior distribution.

Recall that the state of the model consists of both the locations of latent cluster centres and the allocation mapping of the offspring points. We define the state space of the Markov chain as $\{\boldsymbol{\tau}\}$ with $\boldsymbol{\tau} = (\mathbf{x}, \psi_{\mathbf{x}}|\mathbf{u})$, where each state $\boldsymbol{\tau}$ determines a unique configuration of $\{(\boldsymbol{x}_i, \{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(i)}\})\}$ given $\mathbf{u}$. A state $\boldsymbol{\tau}'$ distinct from another $\boldsymbol{\tau}$ can be that it has a different set of cluster centres, or a change in the allocation mapping, or a combination of both. To ensure that the equilibrium distribution of the designed Markov chain aligns with the posterior distribution of the independent cluster model determined by (5.12) and (5.13), we propose four types of move and construct corresponding transition rules that satisfy the detailed balance condition. Although detailed balance imposes stricter constraints than actually needed, it facilitates a more tractable design. For completeness, we will conclude this section by summarizing the designed sampling algorithm and proving the convergence of our MCMC method.

Move type 1: merging and splitting clusters

This move type is similar to the birth and death moves that are commonly used for sampling from a Cox cluster point process (Preston, 1975). Note that, we also need to reallocate the offspring points of the merged or split cluster. Take the merging move as an example. The procedure involves randomly selecting two clusters to merge, deciding the new location of the merged cluster, and combining the offspring points from the two clusters into the new cluster. Denote the centres of the two selected clusters by $\boldsymbol{x}_k, \boldsymbol{x}_l \in \mathbf{x}$, with their corresponding sets of offspring points represented by $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\}, \{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(l)}\}$, and denote the new cluster centre by $\boldsymbol{\xi}$. The state information at this move changes from $\mathbf{x}$ to $\mathbf{x} \setminus \{\boldsymbol{x}_k, \boldsymbol{x}_l\} \cup \{\boldsymbol{\xi}\}$ and the parent of $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\}$ and $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(l)}\}$ becomes $\boldsymbol{\xi}$. Therefore, the posterior probability ratio, recalling (5.12)–(5.14), for the merging move is

$$P_{\text{merge}}(\boldsymbol{x}_k, \boldsymbol{x}_l, \boldsymbol{\xi}|\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = \frac{P\left(\mathbf{x} \setminus \{\boldsymbol{x}_k, \boldsymbol{x}_l\} \cup \{\boldsymbol{\xi}\}\right) g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\right\} \bigcup \left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(l)}\right\} |\boldsymbol{\xi}\right)}{P(\mathbf{x}) g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\right\} |\boldsymbol{x}_k\right) g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(l)}\right\} |\boldsymbol{x}_l\right) \exp(-|\mathcal{S}|)} \tag{5.16}$$

where the additional term $\exp(-|\mathcal{S}|)$ in the denominator accounts for the dimension change. The proposal probability for this move is

$$P^p_{\text{merge}}(\boldsymbol{x}_k, \boldsymbol{x}_l, \boldsymbol{\xi}|\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = \frac{2}{n\{\mathbf{x}\}\left(n\{\mathbf{x}\} - 1\right)} P\left(\boldsymbol{\xi}|\boldsymbol{x}_k, \boldsymbol{x}_l\right), \tag{5.17}$$

where $P(\boldsymbol{\xi}|\boldsymbol{x}_k, \boldsymbol{x}_l)$ is the probability density or mass function of sampling a new location $\boldsymbol{\xi} \in \mathcal{S}^+$. One may simply consider a uniform distribution over $\mathcal{S}^+$ as $P(\boldsymbol{\xi}|\boldsymbol{x}_k, \boldsymbol{x}_l)$, however, to improve sampling efficiency, we set $\boldsymbol{\xi}$ to be either $\boldsymbol{x}_k$ or $\boldsymbol{x}_l$ thus $P(\boldsymbol{\xi}|\boldsymbol{x}_k, \boldsymbol{x}_l) = 1/2$. On the other hand, the reverse proposal probability

is

$$P^{rp}_{\text{merge}}(\boldsymbol{x}_k, \boldsymbol{x}_l, \boldsymbol{\xi}|\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = \frac{P(\boldsymbol{x}_k|\boldsymbol{\xi})P(\boldsymbol{x}_l|\boldsymbol{\xi})}{n\{\mathbf{x}\}-1}\left(\frac{1}{2}\right)^{n\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\right\}}\left(\frac{1}{2}\right)^{n\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(l)}\right\}}. \tag{5.18}$$

Again, for efficiency, we let $\boldsymbol{\xi} = \boldsymbol{x}_k$ thus $P(\boldsymbol{x}_k|\boldsymbol{x}_k) = 1$, and set $P(\boldsymbol{x}_l|\boldsymbol{x}_k)$ to a Gaussian density centred at $\boldsymbol{x}_k$ with a covariance matrix $\boldsymbol{\Sigma}_{\text{P}}$. The last two terms in (5.18) represents the likelihood of evenly reallocating the offspring points between the two clusters. Analogously, one can derive the posterior probability ratio and the proposal and reverse proposal probabilities for the splitting move.

Move type 2: shifting the location of a cluster centre

This move type acts as a shortcut of the first one. Consider the case where a cluster of offspring points are close to each other while the current centre is mis-estimated. Instead of executing multiple merging and splitting moves, i.e. deleting the mis-estimated cluster centre, merging its offspring points into another cluster and splitting them again to obtain a new cluster, the shifting move bypasses the intermediate steps. It directly selects the cluster to shift, relocates its centre, and attaches the offspring points. Denote the centre of the selected cluster by $\boldsymbol{x}_k \in \mathbf{x}$, with its offspring points represented by $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\}$, and denote the new centre location by $\boldsymbol{\xi}$. The state information at this move changes from $\mathbf{x}$ to $\mathbf{x} \setminus \{\boldsymbol{x}_k\} \cup \{\boldsymbol{\xi}\}$ and the parent of $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\}$ becomes $\boldsymbol{\xi}$. The posterior probability ratio, recalling (5.12)–(5.14), for the shifting move is

$$P_{\text{shift}}(\boldsymbol{x}_k, \boldsymbol{\xi}|\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = \frac{P\left(\mathbf{x} \setminus \{\boldsymbol{x}_k\} \cup \{\boldsymbol{\xi}\}\right) g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\right\}|\boldsymbol{\xi}\right)}{P(\mathbf{x}) g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\right\}|\boldsymbol{x}_k\right)}, \tag{5.19}$$

and the proposal probability is $P^{p}_{\text{shift}}(\boldsymbol{x}_k, \boldsymbol{\xi}|\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = P(\boldsymbol{\xi}|\boldsymbol{x}_k)/n\{\mathbf{x}\}$. Here, $P(\boldsymbol{\xi}|\boldsymbol{x}_k)$ denotes the probability density of sampling a new location $\boldsymbol{\xi} \in \mathcal{S}^+$ and is again set to a Gaussian density centred at $\boldsymbol{x}_k$ with a covariance matrix $\boldsymbol{\Sigma}_p$. The reverse proposal probability for this move is $P^{rp}_{\text{shift}}(\boldsymbol{x}_k, \boldsymbol{\xi}|\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = P(\boldsymbol{x}_k|\boldsymbol{\xi})/n\{\mathbf{x}\}$. For convenience, we set $P(\boldsymbol{x}_k|\boldsymbol{\xi})$ to be the same symmetric Gaussian density as $P(\boldsymbol{\xi}|\boldsymbol{x}_k)$ so that $P^{p}_{\text{shift}}(\boldsymbol{x}_k, \boldsymbol{\xi}|\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = P^{rp}_{\text{shift}}(\boldsymbol{x}_k, \boldsymbol{\xi}|\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u})$.

Move type 3: switching the cluster label of an offspring point

This move type modifies a state by adjusting the allocation label of an offspring point while keeping all cluster centres unchanged. The procedure involves randomly selecting an offspring point to switch, reallocating it from the current cluster to a new one, and updating the sets of offspring points for both clusters. Denote the selected offspring point by $\boldsymbol{\omega} \in \mathbf{u}$, and suppose that it is switched from the cluster centred at $\boldsymbol{x}_k$ to the cluster centred at $\boldsymbol{x}_l$. The original sets

of offspring points of the two clusters are $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\}$ and $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(l)}\}$. The related state information at this move changes to $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\} \setminus \{\boldsymbol{\omega}\}$ and $\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(l)}\} \cup \{\boldsymbol{\omega}\}$. Therefore, the posterior probability ratio, recalling (5.12)–(5.14), for the switching move is

$$P_{\text{switch}}(\boldsymbol{\omega}, \boldsymbol{x}_k, \boldsymbol{x}_l | \mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = \frac{g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\right\} \setminus \{\boldsymbol{\omega}\} | \boldsymbol{x}_k\right) g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\right\} \bigcup \{\boldsymbol{\omega}\} | \boldsymbol{x}_l\right)}{g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(k)}\right\} | \boldsymbol{x}_k\right) g\left(\left\{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(l)}\right\} | \boldsymbol{x}_l\right)}, \tag{5.20}$$

and both the proposal and reverse proposal probabilities are

$$\begin{aligned} P^{p}_{\text{switch}}(\boldsymbol{\omega}, \boldsymbol{x}_k, \boldsymbol{x}_l | \mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) &= P^{rp}_{\text{switch}}(\boldsymbol{\omega}, \boldsymbol{x}_k, \boldsymbol{x}_l | \mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) \\ &= \frac{1}{n\{\mathbf{u}\}\left(n\{\mathbf{x}\} - 1\right)}. \end{aligned} \tag{5.21}$$

Move type 4: adding and deleting an empty cluster at the outer boundary

An important practical consideration arises when a cluster centre located outside the observation window of offspring points $\mathcal{S}$, but near the boundary, may produce children within $\mathcal{S}$, as discussed in Section 5.2.2. Conversely, even though no offspring points are observed at the inner boundary of $\mathcal{S}$, there remains a small possibility of cluster centres existing just at the outer boundary of $\mathcal{S}$. This phenomenon is known as 'sampling bias' in Van Lieshout & Baddeley (2002). To account for it, we propose an additional move type that introduces such randomness efficiently by adding and deleting clusters at the outer boundary that appear empty in $\mathcal{S}$. Take the move of adding such an cluster as an example. The procedure involves randomly selecting a location in the area near the outer boundary as the centre of the new cluster and attaching an 'empty' set of offspring points to it. Denote the added cluster centre by $\boldsymbol{\xi}$. The state information at this move changes from $\mathbf{x}$ to $\mathbf{x} \cup \{\boldsymbol{\xi}\}$ and includes a cluster centred at $\boldsymbol{\xi}$. The posterior probability ratio, recalling (5.12)–(5.14), for this move is

$$P_{\text{add}}(\boldsymbol{\xi} | \mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = \frac{P(\mathbf{x} \cup \{\boldsymbol{\xi}\}) g(\emptyset | \boldsymbol{\xi})}{P(\mathbf{x}) \exp(|\mathcal{S}|)}, \tag{5.22}$$

and the proposal probability is $P^{p}_{\text{add}}(\boldsymbol{\xi} | \mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = 1/(|\mathcal{S}^{+} \setminus \mathcal{S}|)$, where $\mathcal{S}^{+} \setminus \mathcal{S}$ represents the area outside $\mathcal{S}$ but within $\mathcal{S}^{+}$. The reverse proposal probability for this move is $P^{rp}_{\text{add}}(\boldsymbol{\xi} | \mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u}) = 1/n\{i : \boldsymbol{x}_i \in \mathbf{x} \cup \{\boldsymbol{\xi}\}, \boldsymbol{x}_i \in \mathcal{S}^{+}, \boldsymbol{x}_i \notin \mathcal{S}, \{\boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(i)}\} = \emptyset\}$, whose denominator counts the number of 'empty' clusters located in the outer boundary area. Similarly, one can derive the posterior probability ratio and the proposal and reverse proposal probabilities for the move of deleting such an 'empty' cluster.

It is important to highlight that, in all four move types, we introduce the posterior probability ratio of the independent cluster model in general terms of

probability densities of the ground process and the mark kernel. Therefore, the proposed MCMC method can be easily adapted to various structures of cluster centres and offspring points by simply substituting those functions with specific expressions.

To sample the states $\boldsymbol{\tau} = (\mathbf{x}, \psi_{\mathbf{x}}, \mathbf{u})$, we consider a standard Metropolis-Hasting algorithm. Let $\boldsymbol{\tau}$ be the current state of the Markov chain. At each iteration, we first generate a uniform random number $d \in [0, 1]$ to decide the next move type to perform. Specifically, if $d < 1/6$, the merging move is selected; if $1/6 \leq d < 1/3$, the splitting move is selected, and so on. With a determined move type, we propose a random candidate state $\boldsymbol{\tau}'$ according to the proposal distribution in that move and compute the acceptance ratio

$$A_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}') = \min\left\{1, \frac{P_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')P^{rp}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')}{P^{p}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')}\right\}. \tag{5.23}$$

For instance, in the merging move, $P_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}'), P^{rp}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}'), P^{p}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')$ are given by (5.16)–(5.18), respectively. Then, we generate another uniform random number $e \in [0, 1]$. If $e \leq A_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')$, we accept $\boldsymbol{\tau}'$ to be the next state; otherwise, the current state $\boldsymbol{\tau}$ stays. The transition probability for the move is thus $P(\boldsymbol{\tau}, \boldsymbol{\tau}') = \min\{P^{p}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}'), P_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')P^{rp}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')\}$. Such an algorithm guarantees the convergence of the proposed MCMC method in the next theorem.

Theorem 5.1. *Let $\{\boldsymbol{\tau}\}_{t=1}^{\infty}$ be the Markov chain defined in this section, equipped with the Metropolis-Hasting algorithm introduced above to sample $\boldsymbol{\tau}$. For almost all states with a positive probability density as the initial state $\boldsymbol{\tau}_0$, the distribution of this Markov chain converges to the posterior distribution of the independent cluster process determined by (5.12) and (5.13) as $t \to \infty$. Moreover, denote*

$$\pi(\boldsymbol{\tau}) = P(\mathbf{x}|\mathbf{u})P(\psi_{\mathbf{x}}|\mathbf{x}, \mathbf{u}),$$

then, for any measurable test function $f : \{\boldsymbol{\tau}\} \to \mathbb{R}$, the MCMC estimate of the expectation of $f(\boldsymbol{\tau})$ converges almost surely to the true expectation under π:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(\boldsymbol{\tau}_t) = \mathbb{E}_{\pi}[f(\boldsymbol{\tau})|\mathbf{u}].$$

The theorem is proved by verifying the detailed balance condition, along with the irreducibility and aperiodicity properties of the Markov chain. The details are given in Appendix 5.8. The convergence result of f allows us to approximate the expectation of test functions based on MCMC samples, which will be used in the following section to construct the expectation-maximization algorithm for estimating the parameters of the independent cluster process model.

5.3.2 Parameter Estimation by Monte Carlo Expectation-maximization

To estimate the intensity and shape parameters of a cluster in Cox cluster processes, Van Lieshout & Baddeley (2002) considered a Monte Carlo maximum like-

lihood method. Similar techniques, equipped with an expectation-maximization algorithm, have been used to estimate parameters in deep Neyman-Scott processes (Hong & Shelton, 2022). For independent cluster process models, we need to estimate the parameters of both the ground process and the mark kernel, simultaneously. To this end, recalling (5.12) and (5.13), we propose to maximize the following integrated likelihood given observed offspring points

$$P(\mathbf{u};\boldsymbol{\theta}) = \int P(\mathbf{x};\boldsymbol{\theta}) \left[\sum_{\psi_{\mathbf{x}}} P(\psi_{\mathbf{x}}, \mathbf{u};\boldsymbol{\theta}|\mathbf{x}) \right] \mathrm{d}\mathbb{P}(\mathbf{x}), \tag{5.24}$$

where we use $\boldsymbol{\theta}$ to represent the set of all parameters in the independent cluster process model and $\mathbb{P}(\mathbf{x})$ to denote the reference probability measure of observing $\mathbf{x}$ from a unit-rate Poisson process on $\mathcal{S}^+$. In the case of (5.2), (5.7) and (5.8), $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \mu, \Sigma\}$. Note that the parameters specifying the interaction range in a Markov point process prior, cf. Section 5.2.1, also need to be approximated, however, we treat them as an nuisance parameter in this chapter. In practice, these parameters can be determined by a profile likelihood approach (Murphy & Van der Vaart, 2000).

Directly maximizing (5.24) is intractable due to the complexity introduced by the latent state variables $(\mathbf{x}, \psi_{\mathbf{x}})$. To address this, we develop an efficient expectation-maximization algorithm (Dempster, Laird & Rubin, 1977) to estimate the parameters $\boldsymbol{\theta}$, iteratively. At each iteration, the integration of the latent state information is approximated by samples obtained using the MCMC method proposed in Section 5.3.1. Specifically, our MCEM algorithm iterates between calculating the expected complete-data log-likelihood given observed offspring points

$$\begin{aligned} &Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = \mathbb{E}_{\boldsymbol{\theta}_t}\left\{\log\left[P(\mathbf{x};\boldsymbol{\theta})P(\psi_{\mathbf{x}}, \mathbf{u};\boldsymbol{\theta}|\mathbf{x})\right]|\mathbf{u}\right\} = \\ &\quad \mathbb{E}_{\boldsymbol{\theta}_t}\left(\log\left\{ P(\mathbf{x};\boldsymbol{\theta}) \prod_{i=1}^{n\{\mathbf{x}\}} \left[g\left(\left\{ \boldsymbol{u}_{\psi_{\mathbf{x}}^{-1}(i)} \right\}; \boldsymbol{\theta}|\boldsymbol{x}_i \right) \exp(-|\mathcal{S}|) \right] \right\} |\mathbf{u} \right) \end{aligned} \tag{5.25}$$

and maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ over $\boldsymbol{\theta}$. Here, $\boldsymbol{\theta}_t$ denotes the maximum of $\boldsymbol{\theta}$ estimated at the t-th iteration in the algorithm. The expectation above is taken over the latent state variables $(\mathbf{x}, \psi_{\mathbf{x}})$ and can be approximated by

$$\frac{1}{J}\sum_{j=1}^{J} \log\left\{ P(\mathbf{x}_j;\boldsymbol{\theta}) \prod_{i=1}^{n\{\mathbf{x}_j\}} \left[g\left(\left\{ \boldsymbol{u}_{\psi_{\mathbf{x}_j}^{-1}(i)} \right\}; \boldsymbol{\theta}|\boldsymbol{x}_i \in \mathbf{x}_j \right) \exp(-|\mathcal{S}|) \right] \right\} \tag{5.26}$$

with $(\mathbf{x}_1, \psi_{\mathbf{x}_1}), \ldots, (\mathbf{x}_J, \psi_{\mathbf{x}_J})$ the samples of states drawn from the posterior distribution of the independent cluster process model given the parameters $\boldsymbol{\theta}_t$.

It is important to note that sampling $(\mathbf{x}_1, \psi_{\mathbf{x}_1}), \ldots, (\mathbf{x}_J, \psi_{\mathbf{x}_J})$ is required at every iteration as $\boldsymbol{\theta}_t$ is constantly updated. However, the computational cost of this procedure can be high. To improve efficiency, we apply importance sampling (Levine & Casella, 2001) to account for new samples by weighting the fixed

samples from a reference posterior distribution. Let $\boldsymbol{\theta}_r$ be the parameters for the reference distribution and denote the state samples drawn according to (5.12) and (5.13) given $\boldsymbol{\theta}_r$ by $(\mathbf{x}_{r,1}, \psi_{\mathbf{x}_{r,1}}), \ldots, (\mathbf{x}_{r,J}, \psi_{\mathbf{x}_{r,J}})$. At the $(t+1)$th-iteration, instead of generating new samples $(\mathbf{x}_1, \psi_{\mathbf{x}_1}), \ldots, (\mathbf{x}_J, \psi_{\mathbf{x}_J})$ based on the estimated parameters $\boldsymbol{\theta}_t$, we importance weight the reference samples $(\mathbf{x}_{r,1}, \psi_{\mathbf{x}_{r,1}}), \ldots, (\mathbf{x}_{r,J}, \psi_{\mathbf{x}_{r,J}})$ and approximate

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) \simeq \sum_{j=1}^{J} w_{t,j} \log \left[P(\mathbf{x}_{r,j}; \boldsymbol{\theta}) \prod_{i=1}^{n\{\mathbf{x}_{r,j}\}} g\left(\left\{ \boldsymbol{u}_{\psi^{-1}_{\mathbf{x}_{r,j}}(i)} \right\}; \boldsymbol{\theta} | \boldsymbol{x}_i \in \mathbf{x}_{r,j} \right) \right] / \sum_{j=1}^{J} w_{t,j}, \tag{5.27}$$

where

$$w_{t,j} = \frac{C(\boldsymbol{\theta}_r|\mathbf{u}) p(\mathbf{x}_{r,j}; \boldsymbol{\theta}_t) \prod_{i=1}^{n\{\mathbf{x}_{r,j}\}} g\left(\left\{ \boldsymbol{v}_{\psi^{-1}_{\mathbf{x}_{r,j}}(i)} \right\}; \boldsymbol{\theta}_t | \boldsymbol{x}_i \in \mathbf{x}_{r,j} \right)}{C(\boldsymbol{\theta}_t|\mathbf{u}) p(\mathbf{x}_{r,j}; \boldsymbol{\theta}_r) \prod_{i=1}^{n\{\mathbf{x}_{r,j}\}} g\left(\left\{ \boldsymbol{v}_{\psi^{-1}_{\mathbf{x}_{r,j}}(i)} \right\}; \boldsymbol{\theta}_r | \boldsymbol{x}_i \in \mathbf{x}_{r,j} \right)}. \tag{5.28}$$

Here, $p(\mathbf{x}; \cdot)$ denotes the unnormalized part in (5.2). Moreover, the constant ratio $C(\boldsymbol{\theta}_r|\mathbf{u})/C(\boldsymbol{\theta}_t|\mathbf{u})$ applies to every weight $w_{t,j}$ with $j = 1, \ldots, J$ and will be cancelled out in (5.27).

If one decides to use a pairwise interaction point process, cf. Section 5.2.1, as the ground process, an additional concern is that $P(\mathbf{x}; \boldsymbol{\theta})$ in (5.27) includes a normalizing constant, denoted as $C(\boldsymbol{\alpha})$ recalling (5.2). This constant does not depend on the latent state variables $(\mathbf{x}, \psi_{\mathbf{x}})$ but changes along with different $\boldsymbol{\theta}$, and needs to be approximated to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$. Again, write $P(\mathbf{x}; \boldsymbol{\alpha}) = C(\boldsymbol{\alpha})^{-1} p(\mathbf{x}; \boldsymbol{\alpha})$ with $p(\mathbf{x}; \boldsymbol{\alpha})$ denoting the unnormalized part in (5.2). Since

$$\frac{C(\boldsymbol{\alpha})}{C(\boldsymbol{\alpha}')} = \int \frac{p(\mathbf{x}; \boldsymbol{\alpha})}{p(\mathbf{x}; \boldsymbol{\alpha}')} \frac{p(\mathbf{x}; \boldsymbol{\alpha}')}{C(\boldsymbol{\alpha}')} \mathrm{d}\mathbb{P}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\alpha}'} \left[\frac{p(\mathbf{x}; \boldsymbol{\theta})}{p(\mathbf{x}; \boldsymbol{\alpha}')} \right], \tag{5.29}$$

we can approximate $C(\boldsymbol{\alpha})$ using another set of samples of $\mathbf{x}$ drawn from the distribution $P(\mathbf{x}; \boldsymbol{\alpha}')$ based on the parameters $\boldsymbol{\theta}'$ as $C(\boldsymbol{\alpha}) = C(\boldsymbol{\alpha}') \sum_{k=1}^{K} \frac{p(\mathbf{x}_k; \boldsymbol{\alpha})}{p(\mathbf{x}_k; \boldsymbol{\alpha}')} / K$. Note that $\mathbf{x}_1, \ldots, \mathbf{x}_K$ are independently generated from $\mathbf{x}_1, \ldots, \mathbf{x}_J$ used in the importance sampling. To ensure the accuracy of this approximation, $\boldsymbol{\alpha}'$ needs to be selected close to $\boldsymbol{\alpha}$. Therefore, at each iteration, we set $\boldsymbol{\theta}' = \boldsymbol{\theta}_t$ and draw new samples $\mathbf{x}_1, \ldots, \mathbf{x}_K$ from the distribution $P(\mathbf{x}; \boldsymbol{\alpha}_t)$.

The importance sampling technique used in (5.27) has the drawback that if the importance density terms are far from the one to be approximated, i.e. the numerator and denominator in $w_{t,j}$, the weights will vary widely, giving many samples little weights and allowing a few variates to be overinfluential. To avoid this phenomenon, we may perform some warm-up iterations with the standard MCEM algorithm to obtain good starting values and conduct the efficient importance sampled algorithm afterwards.

5.4 An Example Model for Arson Fire Data

In this section, we present an example of an independent cluster point process for modelling the arson fire data, which motivates this study, as mentioned in Section 5.1. We start with introducing data patterns and conducting preliminary analyses to guide the model design. Based on the achieved insights, we propose a model structure tailored to capture the specific characteristics of arson clusters and fire incidents.

5.4.1 Arson Fire Data

The data set consists of 326 car arson fire incidents that occurred in the Twente region in the Netherlands from 2017 to 2021. Figure 5.1(left) plots their spatial locations and temporal marks, with larger, darker circles indicating more recent events. Overall, firefighters suggest that arson fires occur in clusters, driven by underlying activity centres of arsonists, where each cluster contains at least one incident. Moreover, the arson centres exhibit spatial and temporal heterogeneity, often appearing in low-income neighbourhoods with a higher demand of social welfare dependence and more frequently on Sundays. For prevention purposes, such as conducting public awareness campaigns to educate potential arsonists and intervening to cease a cluster based on observed incidents, it is necessary to identify historical latent activity centres, detect emerging arson clusters, and predict the fire risk in advance.

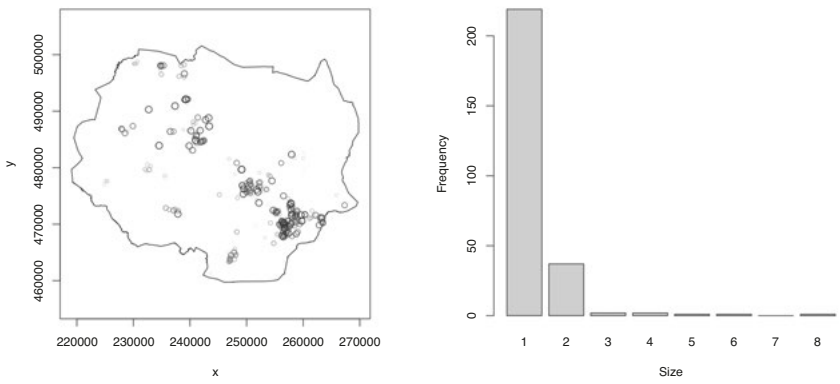


Figure 5.1 Spatio-temporal plot of arson fire incidents (left) and frequencies of cluster sizes identified by a manual analysis (right).

A natural starting point to model the arson fire data is to consider a hierarchical cluster point process, where parent points represent the latent centres of arson activities and offspring points correspond to the observed fire incidents.

However, the two aforementioned challenges discussed in Section 5.1 indicate the need for an independent cluster process model instead of a Neyman-Scott or Cox cluster process model, along with a flexible and robust inference method. Below, we will examine the two challenges in the arson fire data, separately, in detail and discuss their implications for the model design.

Firstly, arson clusters do not adhere to a Poisson structure: the number of arson fires in a cluster is always positive, whereas the Poisson distribution allows for the possibility of empty clusters. To illustrate this, we conduct a manual analysis by linking fire incidents to recorded criminal series in additional data provided by the police, revealing arson clusters with sizes ranging from one to eight fires. Figure 5.1(right) displays the frequencies of various cluster sizes. To capture the distribution of fire intensities in a cluster here, one possible adaptation is to truncate the Poisson distribution by removing the probability mass at zero and normalizing over the remaining values, thereby preserving some computational advantages of the Poisson framework. However, as shown in Figure 5.2(left), the adjusted probability mass function does not fit the data well. A better alternative is to adopt a shifted-Poisson distribution, depicted in Figure 5.2(right), where the probability mass of $a \in \mathbb{N}^+$ is replaced by that of $a-1$ in the original Poisson distribution. While this modification effectively recovers the observed pattern of cluster sizes, it introduces additional complexity to the cluster point process model. Consequently, closed-form expressions for summary statistics become intractable, and the posterior distribution of the model becomes more complicated.

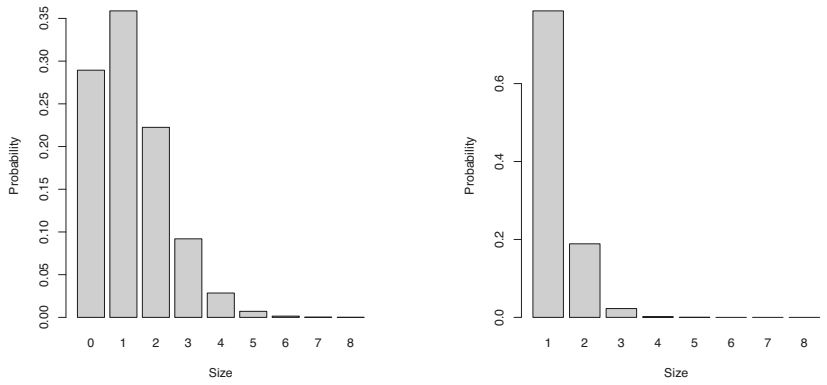


Figure 5.2 Fitted Poisson (left) and shifted Poisson (right) distributions for cluster sizes over the frequencies displayed in Figure 5.1(right).

Secondly, according to police reports, only a small number of arsonists repeatedly commit arson fires. This is also reflected in our manual analysis by the large proportion of single-event clusters, as shown in Figure 5.1(right), indicating weak clustering strength in the fire data. Such a pattern brings challenges to classic in-

ference approaches for producing plausible and stable estimates. To address this, Baddeley et al. (2022) proposed model increments to the Neyman-Scott process within a frequentist framework, however, their method is not directly applicable to general independent cluster processes. An alternative strategy is to treat some of the single-event clusters as background noise, i.e. randomly, individually occurred arson fires, modelled by a Poisson point process, but it introduces the risk of non-identifiability when distinguishing between background noise and events induced by the cluster process (Wang et al., 2023). Our Bayesian inference framework handles this risk by incorporating the uncertainty into the sampling process of the posterior distribution of the independent cluster process.

5.4.2 Model Structure

Based on the investigations above, we design an independent cluster point process customized to model the arson fire data. In line with the framework outlined in Section 5.2, we first introduce the model structure by describing the ground process and the mark kernel. Then, we combine them and derive the terms in analogy to (5.12) and (5.13) that determine the posterior distribution.

For the ground process, we consider a particular type of pairwise interaction point process to model the latent arson activity centres: the inhomogeneous modification of Strauss processes (Strauss, 1975). We assume that its trend function is governed by $\exp[\boldsymbol{\beta}^\top \boldsymbol{z}(\boldsymbol{x})]$ and its interaction function is defined as $\gamma^{1\{\|\boldsymbol{x}_{j,s}-\boldsymbol{x}_{k,s}\|_2<r_s;|\boldsymbol{x}_{j,t}-\boldsymbol{x}_{k,t}|<r_t\}}$, with $\boldsymbol{x}, \boldsymbol{x}_j, \boldsymbol{x}_k \in \mathcal{S}^+$. Here, $\boldsymbol{z}(\cdot)$ and $\boldsymbol{\beta}$ are p-dimensional vectors of covariates and coefficients, both $\in \mathbb{R}^p$, $1\{\cdot\}$ is the indicator function, and $\boldsymbol{x}_{j,s}, \boldsymbol{x}_{k,s}$ and $\boldsymbol{x}_{j,t}, \boldsymbol{x}_{k,t}$ represent spatial and temporal coordinates of the points $\boldsymbol{x}_j, \boldsymbol{x}_k$. The probability density of the ground process with respect to a unit-rate Poisson process on $\mathcal{S}^+$ then becomes

$$P(\mathbf{x}) = \frac{1}{C(\boldsymbol{\beta}, \gamma)} \prod_{i=1}^{n\{\mathbf{x}\}} \exp[\boldsymbol{\beta}^\top \boldsymbol{z}(\boldsymbol{x}_i)] \prod_{1\le j<k\le n\{\mathbf{x}\}} \gamma^{1\{\|\boldsymbol{x}_{j,s}-\boldsymbol{x}_{k,s}\|_2<r_s;|\boldsymbol{x}_{j,t}-\boldsymbol{x}_{k,t}|<r_t\}}. \tag{5.30}$$

Note that this model is well-defined for $0 \le \gamma \le 1$ (Ripley & Kelly, 1977). If $\gamma = 1$, it reduces to an homogeneous Poisson process; if $\gamma = 0$, it is a hard-core process in which no pairs of points lie closer than a distance of (r_s, r_t) in space and time. It again deserves to mention that, by incorporating such a repulsive prior process, the arson fire model will not return excessive clusters of one single fire incident.

For the mark kernel, motivated by Figure 5.2(b), we assume that the number of fires in a arson cluster follows a shifted-Poisson distribution with a mean $\mu > 0$. Therefore, the probability mass function in (5.4) reads $f(a) = \mu^{a-1}e^{-\mu}/(a-1)!$ for $a \in \mathbb{N}^+$. It fulfills the restriction that there must be at least one fire incident

in a cluster. Consequently, referring to (5.6), we have

$$g(\mathbf{y}_i|\boldsymbol{x}_i) = \exp(|S| - \mu)\mu^{n_i-1} \prod_{\boldsymbol{s}\in\mathbf{y}_i} k(\boldsymbol{u}|\boldsymbol{x}_i) \sum_{c_i=0}^{\infty} \frac{c_i + n_i}{c_i!} \left\{\mu\left[1 - q(\boldsymbol{x}_i)\right]\right\}^{c_i}, \quad (5.31)$$

which can be simplified to

$$g(\mathbf{y}_i|\boldsymbol{x}_i) = \exp[|S| - \mu q(\boldsymbol{x}_i)] \frac{n_i + \mu\left[1 - q(\boldsymbol{x}_i)\right]}{\mu} \prod_{\boldsymbol{s}\in\mathbf{y}_i} \mu k(\boldsymbol{s}|\boldsymbol{x}_i). \quad (5.32)$$

Note that (5.31) yields the expression for $n_i \geq 1$ as $a_i \geq 1$. In the special case of $n_i = 0$, where all offspring points fall outside $\mathcal{S}$, the corresponding $g(\mathbf{y}_i|\boldsymbol{x}_i)$ can be derived from

$$\begin{aligned} g(\mathbf{y}_i|\boldsymbol{x}_i) &= \exp(|S|) \sum_{a_i=1}^{\infty} f(a_i) \left[1 - q(\boldsymbol{x}_i)\right]^{a_i} \\ &= \exp[|S| - \mu q(\boldsymbol{x}_i)] \left[1 - q(\boldsymbol{x}_i)\right], \end{aligned} \quad (5.33)$$

which also conforms with the form in (5.32). To model the spread of fire incidents in an arson cluster, we consider a Gaussian scatter allocation kernel of $k(\cdot|\boldsymbol{x}_i)$. We assume that the scatter allocation kernels in all arson clusters share the same covariance matrix and that the kernels in spatial and temporal domains are separable and isotropic. It then yields the following expression

$$\begin{aligned} k(\boldsymbol{s}|\boldsymbol{x}_i) = &\frac{1}{(2\pi)^{3/2}\sigma_s^2\sigma_t} \\ &\exp\left[-\frac{1}{2\sigma_s^2}(\boldsymbol{s}_s - \boldsymbol{x}_{i,s})^\top(\boldsymbol{s}_s - \boldsymbol{x}_{i,s}) - \frac{1}{2\sigma_t^2}(\boldsymbol{s}_t - \boldsymbol{x}_{i,t})^2\right], \end{aligned} \quad (5.34)$$

where σ_s, σ_t denote the spatial and temporal standard deviations, respectively.

Combining the ground process and the mark kernel above, we can derive the conditional probability of observing $\mathbf{u}$ given $\mathbf{x}$, which is given by

$$\begin{aligned} P(\mathbf{u}|\mathbf{x}) = &\exp(|\mathcal{S}|) \exp\left[-\sum_{i=1}^{m} \mu q(\boldsymbol{x}_i)\right] \\ &\sum_{\psi_{\mathbf{x}}} \prod_{i=1}^{m} \left\{ \frac{n\left\{\boldsymbol{v}_{\psi_{\mathbf{x}}^{-1}(i)}\right\} + \mu\left[1 - q(\boldsymbol{x}_i)\right]}{\mu} \prod_{\boldsymbol{u}\in\left\{\boldsymbol{v}_{\psi_{\mathbf{x}}^{-1}(i)}\right\}} \mu k(\boldsymbol{u}|\boldsymbol{x}_i) \right\} \end{aligned} \quad (5.35)$$

Furthermore, for inference purposes under our Bayesian framework, the two terms dominating the posterior distribution of the overall independent cluster model,

cf. (5.12) and (5.13), for the arson fire example, now read

$$P(\mathbf{x}|\mathbf{u}) = \frac{P(\mathbf{u}|\mathbf{x})}{C(\boldsymbol{\beta},\gamma|\mathbf{u})} \prod_{i=1}^{n\{\mathbf{x}\}} \exp[\boldsymbol{\beta}^\top \boldsymbol{z}(\boldsymbol{x}_i)] \prod_{1\le j<k\le n\{\mathbf{x}\}} \gamma^{1\{\|\boldsymbol{x}_{j,s}-\boldsymbol{x}_{k,s}\|_2<r_s;|\boldsymbol{x}_{j,t}-\boldsymbol{x}_{k,t}|<r_t\}} \tag{5.36}$$

and

$$P(\psi_{\mathbf{x}}|\mathbf{x},\mathbf{u}) = \frac{\exp(|\mathcal{S}|)}{P(\mathbf{u}|\mathbf{x})} \exp\left[-\sum_{i=1}^{m} \mu q(\boldsymbol{x}_i)\right] \prod_{i=1}^{m} \left\{ \frac{n\left\{\boldsymbol{v}_{\psi_{\mathbf{x}}^{-1}(i)}\right\} + \mu\left[1-q(\boldsymbol{x}_i)\right]}{\mu} \prod_{\boldsymbol{u}\in\left\{\boldsymbol{v}_{\psi_{\mathbf{x}}^{-1}(i)}\right\}} \mu k(\boldsymbol{u}|\boldsymbol{x}_i) \right\}. \tag{5.37}$$

5.5 Simulation Study

To validate our two-step Bayesian inference approach proposed in Section 5.3, we conduct two experiments on the model structure of an independent cluster process designed for the arson fire data in Section 5.4 to test state and parameter estimation. For visualization purposes, we set the domain of interest to spatial windows $\mathcal{S} \subset \mathcal{S}^+ \subset \mathbb{R}^2$. Accordingly, we reduce (5.30) to

$$P(\mathbf{x}) = \frac{1}{C(\beta,\gamma)} [\exp(\beta)]^{n\{\mathbf{x}\}} \prod_{1\le j<k\le n\{\mathbf{x}\}} \gamma^{1\{\|\boldsymbol{x}_j-\boldsymbol{x}_k\|_2<r\}} \tag{5.38}$$

and (5.34) to

$$k(\boldsymbol{s}|\boldsymbol{x}_i) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{s}-\boldsymbol{x}_i)^\top(\boldsymbol{s}-\boldsymbol{x}_i)\right]. \tag{5.39}$$

To obtain the observed offspring points of the independent cluster model, we first generate a Strauss point process $\mathbf{x}$ on $\mathcal{S}^+$ and produce a shifted-Poisson distributed number of offspring points for each $\boldsymbol{x}_i \in \mathbf{x}$. Then, we allocate these points from $\boldsymbol{x}_i$ to specific locations based on the Gaussian kernel above and collect all offspring points observed within $\mathcal{S}$ as $\mathbf{u}$.

5.5.1 State Estimation

In this experiment, we test the effectiveness of our proposed MCMC method in Section 5.3.1 for sampling the state information of $(\mathbf{x}, \psi_{\mathbf{x}}|\mathbf{x})$ from the posterior distribution of the independent cluster process model, given true values of the parameters of $\beta, \gamma, r, \mu, \sigma$. Specifically, we explore two scenarios of the mark

kernel: one with strong clusters that are well-separated and easy-to-distinguish, i.e. $\mu = 2$ and $\sigma = 0.02$, and the other with weak clusters that are more mixed and sometimes overlapped, i.e. $\mu = 1$ and $\sigma = 0.05$. The former aims to show the fundamental utility of our state estimation method, while the latter evaluates its performance to handle more complex data, e.g. the weakly clustered patterns in arson fire incidents (cf., Section 5.4.1).

For both scenarios, we fix $\mathcal{S}$ to the unit window $[0, 1]^2$ and set $\mathcal{S}^+$ to a slightly larger window which extends 3σ beyond $\mathcal{S}$. The default test values for parameters β, γ, r are chosen to $\log(20), 0.1, 0.1$. For each of these three parameters, we examine an additional value to illustrate how the performance of the proposed method varies when encountering less clustered point patterns. In the implementation, we first generate a data realization of $\mathbf{x}$ and $\mathbf{u}$ for each parameter setting and initialize the Markov chain by assigning each observed offspring point to be an individual cluster, with itself as the cluster centre. We then run the MCMC method for a burn-in period of 100000 steps, followed by the collection of 300 state samples of $(\mathbf{x}, \psi_{\mathbf{x}}|\mathbf{u})$ obtained from the subsequent 150000 steps, with a sampling interval of 500 steps. To visualize the distribution of the sampled cluster centres $\mathbf{x}$, we apply kernel smoothing with a standard deviation of σ. Technically, the evidence of the Markov chain reaching its steady state is reflected in the stabilization of the number of cluster centres in sampled $\mathbf{x}$; an example showing the convergence of our MCMC method is plotted in Figure 5.3.

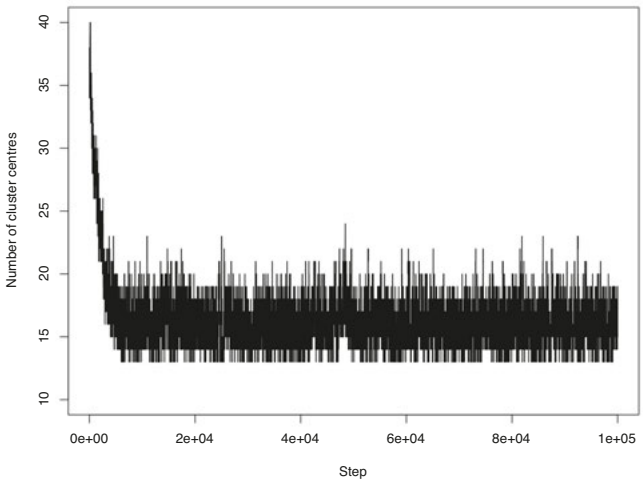


Figure 5.3 The variation in the number of cluster centres within state samples $\mathbf{x}$ when applying the proposed MCMC method to a data realization generated from the independent cluster model with parameters $\beta = \log(20), \gamma = 0.1, r = 0.1, \mu = 2, \sigma = 0.02$ (cf., the first row in Figure 5.4).

The data realizations and the results of the sampled cluster centres $\mathbf{x}$ under different parameter settings in the two scenarios are shown in Figures 5.4 and

5.5, respectively. In the data plots, the black dots represent the generated cluster centres, while the hollow dots denote the offspring points. Overall, in both scenarios, the MCMC samples successfully capture the high-risk regions where latent cluster centres may exist. Since no additional covariate information is involved in the simulation, the distributions of sampled cluster centres are primarily driven by the spatial trends of the observed offspring points. Notably, the state estimation establishes the posterior density surface for the locations of latent cluster centres, which plays a significant role in risk analysis. For instance, in the context of arson fire modelling, these estimates can help identify the neighbourhoods of potential arsonist activity. In addition, comparing Figure 5.4 to Figure 5.5, alongside an analysis on the results under different parameter settings, we observe that when actual cluster strength is weak (i.e., larger β, γ, σ and smaller r, μ), the MCMC estimates of $\mathbf{x}$ also become fused. This behaviour illustrates that, rather than pinpointing precise clusters as in mixture model approaches, our state estimation method emphasizes producing a general risk map. This map highlights the regions where clusters are likely to form, making it particularly suitable for analyzing point pattern data. Furthermore, Figures 5.4 and 5.5 only plot the samples of $\mathbf{x}$. Indeed, the information of $\psi_{\mathbf{x}}$ that specifies the mapping between the sampled cluster centres and the offspring points given $\mathbf{u}$ can also be reached. An example is displayed in Figure 5.6, where colours are added to sampled cluster centres to indicate the attribution of offspring points. Although this mapping is informative, in practice, the interest is more concentrated on $\mathbf{x}$ to provide insights into the distribution of clusters for prediction purposes of new offspring points.

Another advantage of our state estimation method is its capacity to properly recover the risk of latent cluster centres at the boundary region of the observation window of offspring points $\mathcal{S}$. Specifically, three cases can be considered. When a cluster centre locates at the boundary outside $\mathcal{S}$, it can produce offspring points which are observed inside $\mathcal{S}$. The risk associated with this latent cluster centre will be highlighted, resulting in many samples of cluster centres nearby. If none of its offspring points are observed within $\mathcal{S}$, instead of ignoring the existence of this latent cluster centre, a general risk, indicated by random samples of $\boldsymbol{x} \in \mathbf{x}$, as well as the light blue band, along the outer boundary of $\mathcal{S}$, will be added to account for the uncertainty. However, compared to a cluster centre locating within $\mathcal{S}$, this inference for the boundary region is weaker, because for areas well-located in $\mathcal{S}$, if no offspring points are observed nearby, it conveys high confidence that no cluster centre exists in them. Lastly, if a cluster locates at the boundary inside $\mathcal{S}$, it can produce offspring points outside $\mathcal{S}$, which may not be observed. This risk can be captured by random samples of $\boldsymbol{x} \in \mathbf{x}$, as well as the light blue band, along the inner boundary of $\mathcal{S}$.

5.5.2 Parameter Estimation

In the second experiment, we test the MCEM algorithm proposed in Section 5.5.2 for estimating the parameters of the independent cluster process model. Since

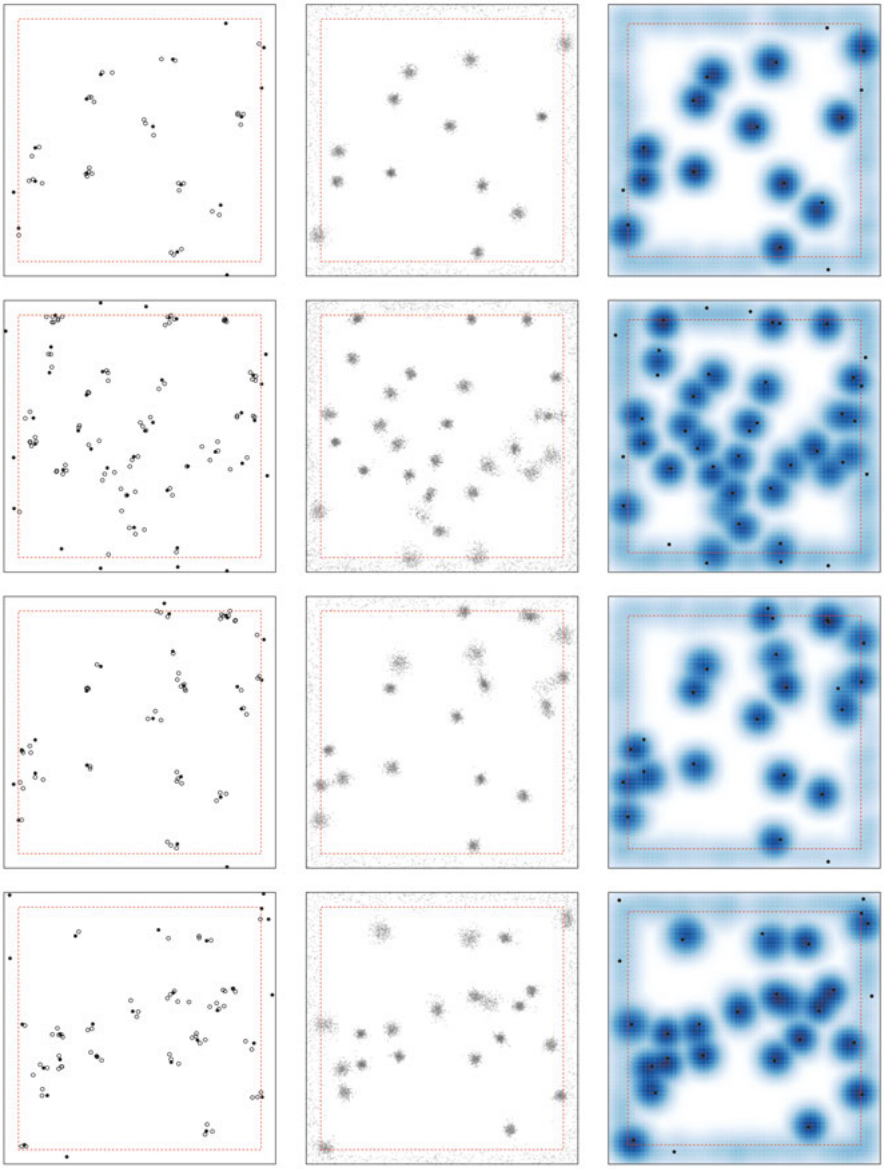


Figure 5.4 Data realizations (left), samples (middle) and smoothed distributions (right) of cluster centres $\mathbf{x}$ obtained by our proposed MCMC method under different parameter settings of β, γ, r when $\mu = 2$ and $\sigma = 0.02$. The first row: $\beta = \log(20), \gamma = 0.1, r = 0.1$; the second row: $\beta = \log(60), \gamma = 0.1, r = 0.1$; the third row: $\beta = \log(20), \gamma = 0.5, r = 0.1$; the forth row: $\beta = \log(20), \gamma = 0.1, r = 0.01$.

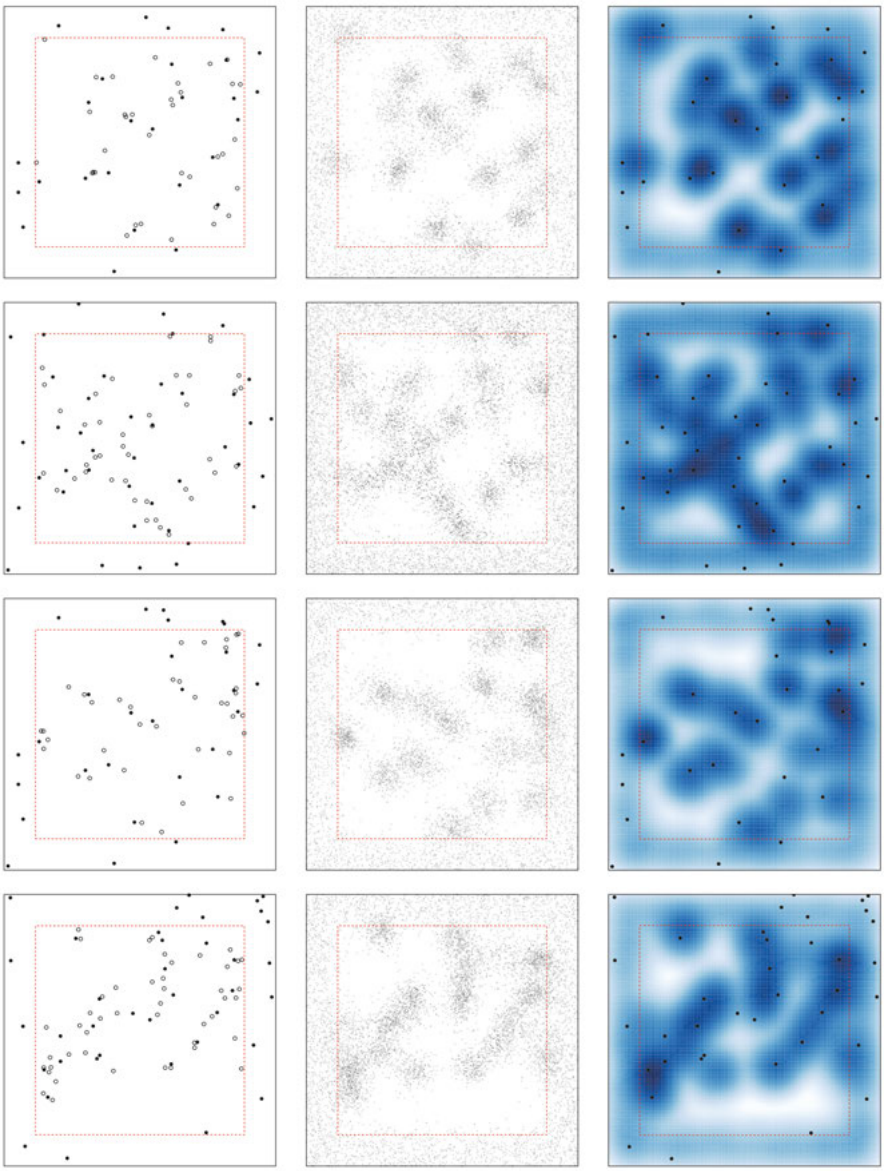


Figure 5.5 Data realizations (left), samples (middle) and smoothed distributions (right) of cluster centres $\mathbf{x}$ obtained by our proposed MCMC method under different parameter settings of β, γ, r when $\mu = 1$ and $\sigma = 0.05$. The first row: $\beta = \log(20), \gamma = 0.1, r = 0.1$; the second row: $\beta = \log(60), \gamma = 0.1, r = 0.1$; the third row: $\beta = \log(20), \gamma = 0.5, r = 0.1$; the fourth row: $\beta = \log(20), \gamma = 0.1, r = 0.01$.

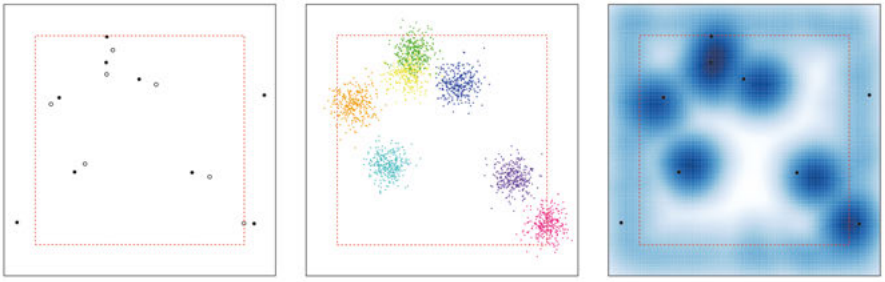


Figure 5.6 An example of data realizations (left), samples (middle) and smoothed distributions (right) of cluster centres $\mathbf{x}$ obtained by the MCMC method, with the mappings between estimated cluster centres and offspring points $\psi_{\mathbf{x}}$ highlighted by different colours.

the interaction range is treated as a nuisance parameter, we fix it to the true value. Again, we test on the two scenarios of the mark kernel with different clustering strengths as introduced in Section 5.5.1. To achieve more stable estimates of the parameters in the Strauss prior (5.36), we set $\mathcal{S} = [0, 1.5]^2$ because there is significant randomness when generating a realization of a Strauss point process on a small window. Accordingly, we set $\mathcal{S}^+$ to a slightly larger window, extending 3σ beyond $[0, 1.5]^2$. The target values for parameters β, γ, r are $\log(20), 0.1, 0.1$. In the implementation, we generate three point patterns of offspring points for each scenario using three different random seeds. We then perform the MCEM algorithm on these data realizations to estimate the parameters $\beta, \gamma, \mu, \sigma$ by conducting 50 warm-up iterations, followed by 50 importance sampling iterations, as detailed in Section 5.3.2. We collect the last 30 iterations of the estimation results of parameters and report the average values in Table 5.1. Moreover, we input these values into the MCMC method to produce the MCMC samples of the state information $\mathbf{x}$ and perform a kernel smoothing again to visualize the estimated high-risk regions where latent cluster centres may exist. The plots of samples and smoothed distributions of cluster centres $\mathbf{x}$ are presented in Figures 5.7 and 5.8.

Table 5.1 Parameter estimates obtained by our proposed MCEM algorithm applied to data realizations generated with different random seeds.

Seed	**True**					**Estimate**			
	β	γ	r	μ	σ	β	γ	μ	σ
1	3	0.1	0.1	2	0.02	3.12	0.107	2.35	0.0214
2	3	0.1	0.1	2	0.02	3.06	0.105	1.82	0.0213
3	3	0.1	0.1	2	0.02	2.78	0.095	1.87	0.0213
1	3	0.1	0.1	1	0.05	3.16	0.103	0.88	0.0393
2	3	0.1	0.1	1	0.05	2.81	0.094	1.36	0.0561
3	3	0.1	0.1	1	0.05	3.19	0.110	1.13	0.0417

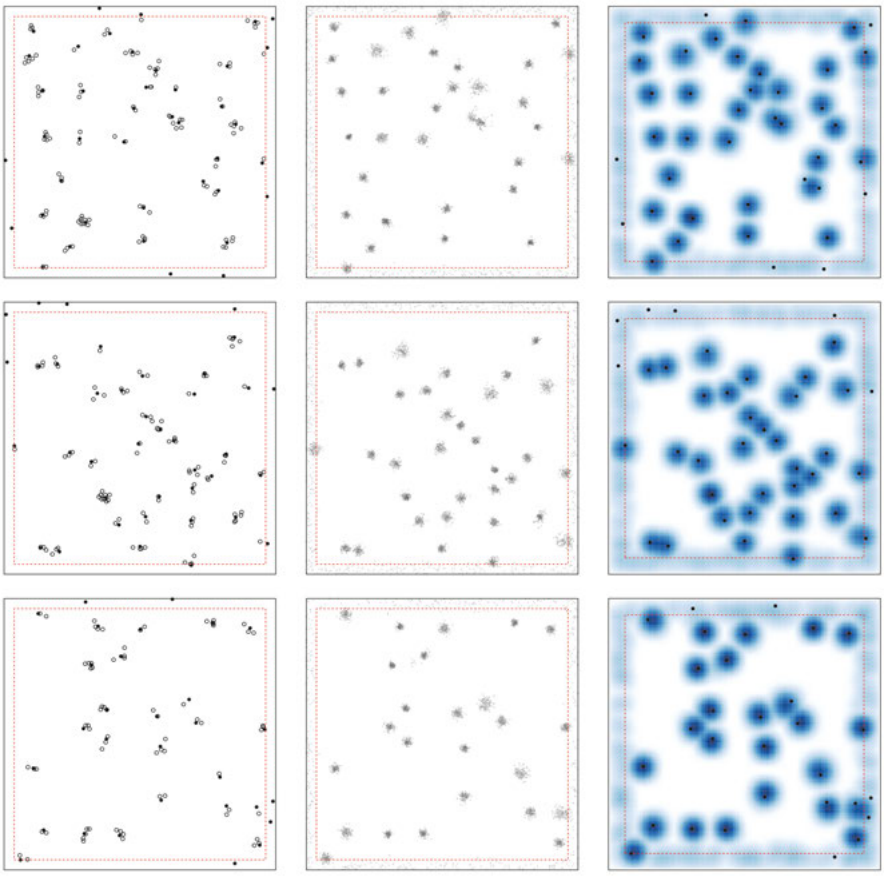


Figure 5.7 Data realizations (left), samples (middle) and smoothed distributions (right) of cluster centres $\mathbf{x}$ obtained by the MCMC method on offspring point patterns generated with different random seeds, using related parameter estimates of $\beta, \gamma, \mu, \sigma$ obtained by the MCEM algorithm, when $\mu = 2$ and $\sigma = 0.02$. The first to third rows correspond to random seeds of 1 to 3.

Overall, the MCEM approach correctly estimates the model parameters of the independent cluster process model, despite some variations arising from the randomness of generating offspring point patterns with different random seeds. The repulsiveness parameter γ in the Strauss prior exhibits relatively low estimation variance due to its sensitivity during data generation. Compared to the clearly clustered scenario (i.e., $\mu = 2$ and $\sigma = 0.02$), parameter estimation is more challenging in the weakly clustered case (i.e., $\mu = 1$ and $\sigma = 0.05$), especially for the shape parameter of a cluster σ. In addition, using the estimated model parameters, the proposed MCMC method effectively recovers high-risk regions of latent cluster centres by sampling the state information $\mathbf{x}$. This includes produc-

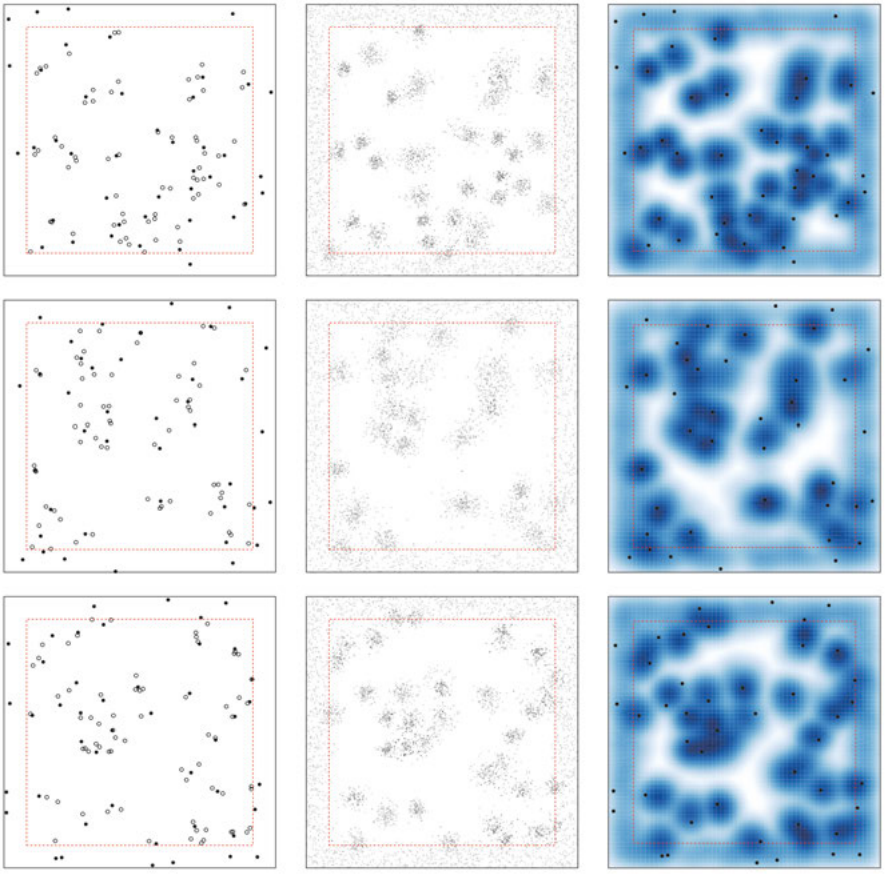


Figure 5.8 Data realizations (left), samples (middle) and smoothed distributions (right) of cluster centres $\mathbf{x}$ obtained by the MCMC method on offspring point patterns generated with different random seeds, using related parameter estimates of $\beta, \gamma, \mu, \sigma$ obtained by the MCEM algorithm, when $\mu = 1$ and $\sigma = 0.05$. The first to third rows correspond to random seeds of 1 to 3.

ing specific risks for the boundary areas, similar to those detailed in Section 5.5.1. The smoothed plot of the sampled states $\mathbf{x}$ by the MCMC method again builds appropriate posterior density surface for the locations of latent cluster centres.

5.6 Application to the Arson Fire Data

As a follow-up step, we plan to infer the proposed independent cluster point process model on the real arson fire data introduced in Section 5.4 by applying the MCMC method and the MCEM algorithm to estimate model parameters and

identify the latent arson activity centres. However, the repulsiveness parameter in the Strauss prior and the shape parameter that determines the spread of offspring points within a cluster are highly sensitive. Selecting reasonable initial values for these parameters poses a challenge, particularly in a spatio-temporal domain of interest. To address, we will need to conduct additional preliminary analyses, incorporating the expertise and experience of firefighters.

At the final stage, the estimated results of the independent cluster process model for arson fire data can be practically utilized in two directions: (i) with the estimated most recent state information, one may extend the MCMC sampling procedure to generate the risk map for the upcoming period, helping fire prediction efforts; (ii) with the estimated intensity parameter μ of an arson cluster, one can perform a threshold analysis to identify potential emerging arson clusters, enabling fire prevention measures.

5.7 Summary

In this chapter, we proposed a two-step Bayesian inference framework to estimate hierarchical independent cluster point processes whose parent and offspring processes can be arbitrarily non-Poisson. It was motivated by a case study on arson fire risk prediction, where arson fire incidents occur in clusters of arsonist activities and is modelled by a independent cluster processes with a Strauss prior for cluster centres and a shifted-Poisson number of offsprings within each cluster. For inference, we develop a Markov chain Monte Carlo method to estimate the state information of the complete process given fixed model parameters and design a Monte Carlo expectation-maximization algorithm to estimate the parameters with obtained MCMC samples. We validate the effectiveness of the proposed approach through a simulation study and will apply it to the real fire data to demonstrate its practical utility. Our Bayesian framework offers significant flexibility for different ground processes and mark kernels which can introduce various interactions between cluster centres and accommodate customized cluster structures provided that their probability functions have an analytical form.

For future work, note that the independent cluster process still assumes independence among the generation of clusters given cluster centres. It would also be interesting to study more general processes where the realization of an offspring point in a cluster not only depends on its own cluster centre but also on offspring points from neighbouring clusters.

5.8 Appendix

Proof of Theorem 5.1

The verification of the detailed balance condition is straightforward, due to the standard Metropolis-Hasting algorithm we use. Note that the transition proba-

bility in a move is $P(\boldsymbol{\tau}, \boldsymbol{\tau}') = \min\{P^p_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}'), P_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')P^{rp}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')\}$, then

$$\begin{aligned}\pi(\boldsymbol{\tau})P(\boldsymbol{\tau}, \boldsymbol{\tau}') &= \pi(\boldsymbol{\tau}) \min\left\{P^p_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}'), \frac{\pi(\boldsymbol{\tau}')}{\pi(\boldsymbol{\tau})}P^{rp}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')\right\} \\ &= \min\left\{\pi(\boldsymbol{\tau})P^p_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}'), \pi(\boldsymbol{\tau}')P^{rp}_{\text{move}}(\boldsymbol{\tau}, \boldsymbol{\tau}')\right\} \\ &= \pi(\boldsymbol{\tau}')P(\boldsymbol{\tau}', \boldsymbol{\tau}).\end{aligned}$$

Irreducibility is guaranteed by the fact that both the proposal and reverse proposal probabilities in our move designs are always positive. For any two states $\boldsymbol{\tau}_i, \boldsymbol{\tau}_j$ in the state space of our Markov chain, where $\pi(\boldsymbol{\tau}_i), \pi(\boldsymbol{\tau}_j) > 0$, one can transition to the other in a finite number of steps by adding or deleting the clusters and reallocating the labels of the offspring points. The transition probability is positive as long as, for all intermediate states in $\boldsymbol{\tau}_i, \ldots, \boldsymbol{\tau}_t, \ldots, \boldsymbol{\tau}_j$, $\pi(\boldsymbol{\tau}_t) > 0$. Exceptions may occur, for example, when $\gamma = 0$ in (5.30), $P(\mathbf{x}) = 0$ if two cluster centres are closer than the interaction range (r_s, r_t), resulting in $\pi(\boldsymbol{\tau}) = 0$. However, our Metropolis-Hasting algorithm avoids moving into such states as long as the initial state satisfies $\pi(\boldsymbol{\tau}) > 0$, because the transition probability into these states would be zero. Aperiodicity can be verified by $P(\boldsymbol{\tau}, \boldsymbol{\tau}) > 0$, which is again straightforward since the proposal and reverse proposal probabilities are positive. The proof is then completed by applying the strong law of large numbers to the Markov chain when computing the expectation of a test function f.

CHAPTER 6

Conclusions and Future Directions

6.1 Conclusions

This thesis addressed four important research challenges within the field of spatial and spatio-temporal point processes: variable selection in point process modelling, uncertainty quantification under the infill regime, flexible nonparametric intensity estimation, and Bayesian inference for hierarchical independent cluster models. These challenges were motivated by fire risk prediction problems in the Netherlands. While some of these issues have received limited attention in traditional statistical point process literature, they are of great significance for effectively modelling and analyzing point pattern data in real-world applications.

To manage an extensive set of environmental variables when building a point process model, we developed a combined machine learning and statistical modelling procedure and applied it to a case study on chimney fire risk prediction. We first used random forests and permutation importance techniques to identify the most informative explanatory variables and then designed a Poisson point process model and employed logistic regression estimation to estimate the parameters. We further validated the Poisson model assumption using second-order summary statistics and residuals. By implementing this modelling procedure on data collected by the Twente Fire Brigade, we obtained plausible fire risk predictions. Compared to similar studies, our approach has two advantages. With random forests, we can select explanatory variables nonparametrically considering variable dependence. Using logistic regression estimation, we can fit the statistical model efficiently by tuning it to focus on important regions and times.

To construct confidence intervals for model estimates in chimney fire prediction under the infill regime, we derived asymptotic results for independent and identically distributed spatio-temporal point processes whose intensity functions are of log-linear form. We first proved strong consistency and asymptotic normality for the parameters of a Poisson point process model estimated by logistic regression estimators and then extended the results to general point process models for which replicated realizations are available. Consistent estimators of the asymptotic covariance matrix were proposed for both cases, and through a simulation study, we demonstrated the effectiveness of a regular dummy point process on variance reduction in logistic regression estimation. Additionally, un-

der proper conditions, we also extended our central limit theorem to general unbiased estimating equations that are based on the Campbell–Mecke theorem. It lays the foundation for further research on the design of test functions in these equations.

To model complex relationships between covariates and the intensity function of a point process and estimate the latter simultaneously, we proposed a nonparametric tree-based ensemble method, which extends the use of gradient-boosted regression trees to the point process context via two carefully designed loss functions. The first loss was based on the Poisson likelihood and works for general point processes. The second loss was derived from a weighted likelihood, where spatially dependent weights are dynamically computed and incorporated to improve the estimation accuracy for clustered point processes. An efficient learning algorithm and an associated validation procedure were designed for model estimation, and the effectiveness of the proposed method was demonstrated through extensive simulation studies and two real data analyses. Notably, XGBoostPP demonstrated strong performance when applied to kitchen fire risk data whose dimension of covariate space is large and training data is very limited, while other approaches encountered computational problems and produced significantly less accurate estimates. It highlights the advantages of using tree ensembles to estimate complex point process intensities.

To estimate the hierarchical independent cluster point process model with arbitrary non-Poisson ground processes and mark kernels, we developed a Bayesian inference framework, motivated by a case study on arson fire risk prediction, where a repulsive Strauss process was employed to model latent cluster centres of arsonist activity and a shifted-Poisson offspring process with spatial and temporal Gaussian spreading kernels was considered for generating the actual fire incidents within each cluster. For inference, we developed a Markov chain Monte Carlo method to estimate cluster states given fixed model parameters and a Monte Carlo expectation-maximization algorithm to estimate the parameters themselves. We showed convergence of the proposed approach and validated its effectiveness through a simulation study on the model structure designed for arson fire data. Our Bayesian framework is flexible, accommodating a wide range of ground processes and mark kernels as long as their probability density functions with respect to related Poisson processes are sufficiently expressive. As a follow-up work, we will apply the proposed inference approach to real arson fire data, identifying underlying arsonist activity centres and predicting emerging arson clusters, which will help provide important insights for risk assessment for firefighters.

In conclusion, the studies presented in this thesis provide both practical and theoretical advancements. Practically, they offer reliable, accurate risk predictions for different fire patterns, with complete confidence intervals, which can be directly applied by the fire service to enhance operational planning and response. Theoretically, the contributions fill in key gaps in the literature on point process modelling and, more importantly, bridge traditional statistical and modern

machine learning communities together.

6.2 Future Directions

Machine and deep learning techniques have become increasingly popular in point pattern analysis due to their strong ability to capture complex data relationships. In addition to the tree-based methods proposed in this thesis for intensity estimation, neural temporal point processes – developed under the framework of Hawkes processes (Hawkes, 1971) – leverage sequential learning architectures of deep neural networks to predict future events based on historical data by estimating conditional intensities (Shchur et al., 2021). Building upon these works, further research could continue to explore the utility of machine and deep learning techniques in the point process domain. For instance, one promising direction would be to investigate covariate-dependent second-order properties of point processes, where interactions of two points depend not only on the distance in between but also on covariate information at the two locations of interest. This concept has been preliminarily studied using parametric models (Tang, 2020), but machine learning methods could enable more flexible, integrated estimators by combining the first- and second-order analyses of point processes into a multiclass classification or regression problem based on the composite likelihood. The challenge, however, lies in finding appropriate machine learning architectures to accommodate particular spatial and spatio-temporal dependence structure inherent to point pattern data.

Secondly, while numerous machine and deep learning approaches have demonstrated superior performance over traditional statistical approaches in point process modelling, establishing theoretical foundations and interpretations is essential to substantiate and understand these advantages fully. Another important research direction, therefore, is to frame machine and deep learning methods within the general framework of statistical models by drawing connections to classic parametric or non-parametric approaches. For instance, in change point detection, deep learning models were formulated as a generalized form of the standard cumulative sum-based classifiers and thus obtained at least as good performance as the latter (Li et al., 2024). We considered something similar for XGBoostPP in Chapter 3, where we interpreted it as a generalized linear model with transformed covariates that denote the membership identities of a point with respect to the supporting leaves of tree ensembles. Beyond interpretability, it is also necessary to examine the estimation efficiency and algorithm complexity of machine and deep learning approaches. Specifically, the research question would be, given a number of data points or replicated point patterns, what are the convergence rates and error bounds for these methods as they approach optimal solutions? To address it, we could adapt the constructed asymptotic and non-asymptotic theories (e.g., Chi et al., 2022; Wang & Lin, 2023) to the context of spatial and spatio-temporal stochastic processes.

The development of methodology and theory for these machine and deep

learning approaches forms a new, important branch in point process modelling. It enables more effective integration of classic spatial and spatio-temporal frameworks with the intricate structures presented in modern, complex point pattern data and deserves further investigation.

Bibliography

Adams, R. P., Murray, I. & MacKay, D. J. C. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th International Conference on Machine Learning*, 9–16. https://dl.acm.org/doi/10.1145/1553374.1553376.

Aglietti, V., Bonilla, E. V., Damoulas, T. & Cripps, S. (2019). Structured variational inference in continuous Cox process models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 12458–12468. https://dl.acm.org/doi/10.5555/3454287.3455403.

Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

Baddeley, A. & Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Australian & New Zealand Journal of Statistics*, 42(3), 283–322. https://doi.org/10.1111/1467-842X.00128.

Baddeley, A. J., Møller, J. & Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3), 329–350. https://doi.org/10.1111/1467-9574.00144.

Baddeley, A., Turner, R., Møller, J. & Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5), 617–666. https://doi.org/10.1111/j.1467-9868.2005.00519.x.

Baddeley, A., Chang, Y.-M., Song, Y. & Turner, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and Its Interface*, 5(2), 221–236. https://intlpress.com/site/pub/files/_fulltext/journals/sii/2012/0005/0002/SII-2012-0005-0002-a007.pdf.

Baddeley, A., Coeurjolly, J.-F., Rubak, E. & Waagepetersen, R. (2014). Logistic regression for spatial Gibbs point processes. *Biometrika*, 101(2), 377–392. https://doi.org/10.1093/biomet/ast060.

Baddeley, A., Rubak, E. & Turner R. (2015). *Spatial Point Patterns: Methodology and Applications with R*, CRC Press. https://doi.org/10.1201/b19708.

Baddeley, A. and Davies, T. M. and Hazelton, M. L. and Rakshit, S. and Turner, R. (2022). Fundamental problems in fitting spatial cluster process models. *Spatial Statistics*, 52, 100709. https://doi.org/10.1016/j.spasta.2022.100709.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis of Spatial Data* (2nd ed.), CRC Press. https://doi.org/10.1201/b17115.

Beraha, M. and Argiento, R. and Møller, J. and Guglielmi, A. (2022). MCMC computations for Bayesian mixture models using repulsive point processes. *Journal of Computational and Graphical Statistics*, 31(2), 422–435. https://doi.org/10.1080/10618600.2021.2000424.

Berman, M. & Turner, T. R. (1992). Approximating Point Process Likelihoods with GLIM. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1), 31–38. https://doi.org/10.2307/2347614.

Besag, J. (1978). Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute*, 47, 77–92.

Boubeta, M., Lombardía, M. J., Marey-Pérez, M. F. & Morales, D. (2015). Prediction of forest fires with area-level Poisson mixed models. *Journal of Environmental Management*, 154, 151–158. https://doi.org/10.1016/j.jenvman.2015.02.009.

Bray, A. & Schoenberg, F. P. (2013). Assessment of point process models for earthquake forecasting. *Statistical Science*, 28(4), 510–520. https://doi.org/10.1214/13-STS440.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785.

Chi, C.-M., Vossler, P., Fan, Y. & Lv, J. (2022). Asymptotic properties of high-dimensional random forests. *Annals of Statistics*, 50(6), 3415–3438. https://doi.org/10.1214/22-AOS2234.

Choiruddin, A., Coeurjolly, J.-F. & Letué, F. (2018). Convex and non-convex regularization methods for spatial point processes intensity estimation. *Electronic Journal of Statistics*, 12(1), 1210–1255. https://doi.org/10.1214/18-EJS1408.

Choiruddin, A., Coeurjolly, J.-F. & Waagepetersen, R. (2020). Information criteria for inhomogeneous spatial point processes. *Australian & New Zealand Journal of Statistics*, 63(1), 119–143. https://doi.org/10.1111/anzs.12327.

Choiruddin, A., Coeurjolly, J.-F. & Letué, F. (2021). Adaptive lasso and Dantzig selector for spatial point processes intensity estimation. *Bernoulli*, 29(3), 1849–1876. https://doi.org/10.3150/22-BEJ1523.

Chiu, S. N., Stoyan, D., Kendall, W. S. & Mecke, J. (2013). *Stochastic Geometry and its Applications*, Wiley. https://doi.org/10.1002/9781118658222.

Cleveland, W. S., Grosse, E. & Shyu, W. M. (1992). Local Regression Models. In *Statistical Models in S* (1st ed.), 8, CRC Press. https://doi.org/10.1201/9780203738535.

Coeurjolly, J.-F. & Guan, Y. (2014). Covariance of empirical functionals for inhomogeneous spatial point processes when the intensity has a parametric form. *Journal of Statistical Planning and Inference*, 155, 79–92. https://doi.org/10.1016/j.jspi.2014.07.003.

Coeurjolly, J.-F. & Møller J. (2014). Variational approach for spatial point process intensity estimation. *Bernoulli*, 20(3), 1097–1125. https://doi.org/10.3150/13-BEJ516.

Coles, P. & Jones, B. (1991). A lognormal model for the cosmological mass distribution. *Monthly Notices of the Royal Astronomical Society*, 248(1), 1–13. https://doi.org/10.1093/mnras/248.1.1.

Condit R., Perez, R., Aguilar, S., Lao, S., Foster, R. & Hubbell, S. P. (2019). Complete data from the Barro Colorado 50-ha plot: 423617 trees, 35 years, 2019 version. https://doi.org/10.15146/5xcp-0d46.

Costafreda-Aumedes, S., Comas, C. & Vega-Garcia, C. (2016). Spatio-temporal configurations of human-caused fires in Spain through point patterns. *Forests*, 7(9), 185. https://doi.org/10.3390/f7090185.

Cronie, O. & Lieshout, M. N. M. van (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105(2), 455–462. https://doi.org/10.1093/biomet/asy001.

Cronie, O., Moradi, M. & Biscio, C. A. N. (2023). A cross-validation-based statistical theory for point processes. *Biometrika*, 111(2), 625–641. https://doi.org/10.1093/biomet/asad041.

Cunningham, J. P., Shenoy, K. V. & Sahani, M. (2008). Fast Gaussian process methods for point process intensity estimation. In *Proceedings of the 25th International Conference on Machine Learning*, 192–199. https://doi.org/10.1145/1390156.1390181.

Daley, D. J. & Vere-Jones, D. (2009). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods* (2nd ed.), Springer. https://doi.org/10.1007/b97277.

D'Angelo, N., Adelfio, G., Abbruzzo, A. & Mateu J. (2022). Inhomogeneous spatio-temporal point processes on linear networks for visitors' stops data. *Annals of Applied Statistics*, 16(2), 791–815. https://doi.org/10.1214/21-AOAS1519.

Debeer, D. & Strobl, C. (2020). Conditional permutation importance revisited. *BMC Bioinformatics*, 21, 307. https://doi.org/10.1186/s12859-020-03622-2.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2), 138–147. https://doi.org/10.2307/2347366.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns* (3rd ed.), CRC Press. https://doi.org/10.1201/b15326.

Dong, Z., Zhu, S., Xie, Y., Mateu, J. & Rodríguez-Cortés F. J. (2023). Non-stationary spatio-temporal point process modeling for high-resolution COVID-19 data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 72(2), 368–386. https://doi.org/10.1093/jrsssc/qlad013.

Donner, C. & Opper, M. (2018). Efficient Bayesian inference of sigmoidal Gaussian Cox processes. *Journal of Machine Learning Research*, 19(1), 2710–2743. https://dl.acm.org/doi/10.5555/3291125.3309629.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M. & Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1555–1564. https://doi.org/10.1145/2939672.2939875.

Everitt, B. S. & Hand, D. J. (1981). *Finite Mixture Distributions*, Springer. https://doi.org/10.1007/978-94-009-5897-5.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics*, 13(1), 342–368. https://doi.org/10.1214/aos/1176346597.

Fiksel, T. (1984). Estimation of parameterized pair potentials of marked and non marked Gibbsian point processes. *Elektronische Informationsverarbeitung und Kybernetik*, 20, 270–278.

Fiksel, T. (1988). Estimation of interaction potentials of Gibbsian point processes. *Statistics*, 19(1), 77–86. https://doi.org/10.1080/02331888808802074.

Flaxman, S., Teh, Y. W. & Sejdinovic, D. (2017). Poisson intensity estimation with reproducing kernels. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 270–279. https://proceedings.mlr.press/v54/flaxman17a.html.

Gabriel, E. & Diggle, P. J. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, 63(1), 43–51. https://doi.org/10.1111/j.1467-9574.2008.00407.x.

Gatrell, A. C., Bailey, T. C., Diggle, P. J. & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21(1), 256–274. https://www.jstor.org/stable/pdf/622936.

Gautier, G., Bardenet, R. & Valko, M. (2019). On two ways to use determinantal point processes for Monte Carlo integration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 7770–7779. https://dl.acm.org/doi/10.5555/3454287.3454985.

Georgii, H. (2011). *Gibbs Measures and Phase Transitions*, De Gruyter. https://doi.org/10.1515/9783110250329.

Godambe, V. P. & Heyde, C. C. (2010). Quasi-likelihood and optimal estimation. *International Statistical Review*, 55(3), 231–244. https://doi.org/10.2307/1403403.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4), 711–732. https://doi.org/10.1093/biomet/82.4.711.

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*, 6, Oxford University Press. https://doi.org/10.1093/oso/9780198510550.001.0001.

Grinsztajn, L., Oyallon, E. & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 507–520. https://dl.acm.org/doi/10.5555/3600270.3600307.

Guan, Y (2006). A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, 101(476), 1502–1512. https://doi.org/10.1198/016214506000000500.

Guan, Y. & Loh, J. M. (2007). A thinned block bootstrap variance estimation procedure for inhomogeneous spatial point patterns. *Journal of the American Statistical Association*, 102(480), 1377–1386. https://doi.org/10.1198/016214507000000879.

Guan, Y (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, 103(483), 1238–1247. https://doi.org/10.1198/016214508000000526.

Guan, Y. & Shen, Y. (2010). A weighted estimating equation approach for inhomogeneous spatial point processes. *Biometrika*, 97(4), 867–880. https://doi.org/10.1093/biomet/asq043.

Guan, Y. & Wang, H. (2010). Sufficient dimension reduction for spatial point processes directed by Gaussian random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3), 367–387. https://doi.org/10.1111/j.1467-9868.2010.00738.x.

Guan, Y., Jalilian, A. & Waagepetersen R. (2015). Quasi-likelihood for spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3), 677–697. https://doi.org/10.1111/rssb.12083.

Gunter, T., Lloyd, C., Osborne, M. A. & Roberts, S. J. (2014). Efficient Bayesian nonparametric modelling of structured point processes. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 310–319. https://dl.acm.org/doi/10.5555/3020751.3020784.

Guyon, X. (1995). *Random Fields on a Network: Modeling, Statistics and Applications*, Springer.

Hackbusch, W. (1995). *Integral Equations: Theory and Numerical Treatment*, Birkhauser Verlag. https://doi.org/10.1007/978-3-0348-9215-5.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1), 83–90. https://doi.org/10.1093/biomet/58.1.83.

Hering, A. S., Bell, C. L. & Genton, M. G. (2009). Modeling spatio-temporal wildfire ignition point patterns. *Environmental and Ecological Statistics*, 16(2), 225–250. https://doi.org/10.1007/s10651-007-0080-6.

Hessellund, K. B. and Xu, G. and Guan, Y. & Waagepetersen, R. (2022). Semiparametric multinomial logistic regression for multivariate point pattern data. *Journal of the American Statistical Association*, 117(539), 1500–1515. https://doi.org/10.1080/01621459.2020.1863812.

Hong, C. & Shelton, C. (2022). Deep Neyman-Scott processes. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 3267–3646. https://proceedings.mlr.press/v151/hong22a.html.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674. https://doi.org/10.1198/106186006X133933.

Hothorn, T., Büehlmann, P., Dudoit, S., Molinaro, A. & Laan, M. J. van der (2006). Survival ensembles. *Biostatistics*, 7(3), 355–373. https://doi.org/10.1093/biostatistics/kxj011.

Illian, J. B., Sørbye, S. H. & Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Annals of Applied Statistics*, 6(4), 1499–1530. https://doi.org/10.1214/11-AOAS530.

Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S.W. & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4), 478–505. https://doi.org/10.1139/er-2020-0019.

John, S. T. & Hensman, J. (2018). Large-scale Cox process inference using variational Fourier features. In *Proceedings of the 35th International Conference on Machine Learning*, 2362–2370. https://proceedings.mlr.press/v80/john18a.html.

Kasteren, L. N. van (2024). Chimney fire prediction in IJsselland. *University of Twente, Bachelor Thesis*. http://essay.utwente.nl/98250.

Kim, H., Asami, T. & Toda, H. (2022). Fast Bayesian estimation of point process intensity as function of covariates. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 25711–25724. https://dl.acm.org/doi/10.5555/3600270.3602134.

Koh, J., Pimont, F., Dupuy, J.-L. & Opitz, T. (2023). Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. *Annals of Applied Statistics*, 17(1), 560–582. https://doi.org/10.1214/22-AOAS1642.

Kutoyants, Y.A. (1998). *Statistical Inference for Spatial Poisson Processes*, Springer. https://doi.org/10.1007/978-1-4612-1706-0.

Lavancier, F., Møller, J. & Rubak, E. (2015). Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 77(4), 853–877. https://doi.org/10.1111/rssb.12096.

Leeuwen, D. van (2022). Data-driven kitchen fire prediction based on environmental variables. *University of Twente, Bachelor Thesis*. http://essay.utwente.nl/92010.

Levine, R. A. & Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3), 422–439. https://doi.org/10.1198/106186001317115045.

Li, J., Fearnhead, P., Fryzlewicz, P. & Wang, T. (2024). Automatic change-point detection in time series via deep learning. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 86(2), 273–285. https://doi.org/10.1093/jrsssb/qkae004

Lieshout, M. N. M. van (2000). *Stochastic geometry models in image analysis and spatial statistics*, Free University Amsterdam, PhD Thesis. https://ir.cwi.nl/pub/13130.

Lieshout, M. N. M. van (2000). *Markov point processes and their applications*, Imperial College Press. https://doi.org/10.1142/p060.

Lieshout, M. N. M. van (2021). Infill asymptotics for adaptive kernel estimators of spatial intensity. *Australian & New Zealand Journal of Statistics*, 63(1), 159–181. https://doi.org/10.1111/anzs.12319.

Lieshout, M. N. M. van & Baddeley, A. J. (2002). Extrapolating and Interpolating Spatial Patterns. In *Spatial Cluster Modelling*, CRC Press. https://doi.org/10.1201/9781420035414.

Lloyd, C., Gunter, T., Osborne, M. A. & Roberts, S. J. (2015). Variational inference for Gaussian process modulated Poisson processes. In *Proceedings of the 32th International Conference on Machine Learning*, 1814–1822. https://dl.acm.org/doi/10.5555/3045118.3045311.

Malik, A., Rao, M. R., Puppala, N., Koouri, P., Anil, V., Thota, V. A. K., Liu, Q., Chiao, S. & Gao, J. (2014). Data-driven wildfire risk prediction in Northern California. *Atmosphere*, 12(1), 109. https://doi.org/10.3390/atmos12010109.

Mandallaz, D. & Ye, R. (2011). Prediction of forest fires with Poisson models. *Canadian Journal of Forest Research*, 27(10), 1685–1694. https://doi.org/10.1139/x97-103.

McCullagh, P. & Møller, J. (2006). The permanental process. *Advances in Applied Probability*, 38(4), 873–888. https://www.jstor.org/stable/20443479.

McCullagh, P. & Nelder, J. A. (2019). *Generalized Linear Models* (2nd ed.), CRC Press. https://doi.org/10.1201/9780203753736.

Mei, H. & Eisner, J. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6757–6767. https://dl.acm.org/doi/10.5555/3295222.3295420.

Miller, J. W. & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521), 340–356. https://doi.org/10.1080/01621459.2016.1255636.

Møller, J., Syversveen, A. R. & Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3), 451–482. https://doi.org/10.1111/1467-9469.00115.

Møller, J. & Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*, Taylor & Francis.

Møller, J. & Torrisi, G. L. (2005). Generalised shot noise Cox processes. *Advances in Applied Probability*, 37(1), 48–74. https://www.jstor.org/stable/30037315.

Møller, J. & Waagepetersen, R. P. (2007). Modern Statistics for Spatial Point Processes. *Scandinavian Journal of Statistics*, 34(4), 643–684. https://doi.org/10.1111/j.1467-9469.2007.00569.x.

Møller J. & Díaz-Avalos C. (2010). Structured spatio-temporal shot-noise Cox point process models, with a view to modelling forest fires. *Scandinavian Journal of Statistics*, 37(1), 2–25. https://doi.org/10.1111/j.1467-9469.2009.00670.x.

Mrkvička, T., Muška, M. & Kubečka, J. (2012). Two step estimation for Neyman-Scott point process with inhomogeneous cluster centers. *Statistics and Computing*, 24, 91–100. https://doi.org/10.1007/s11222-012-9355-3.

Murphy, S. A. & Vaart, A. W. van der (2000). On Profile Likelihood. *Journal of the American Statistical Association*, 95(450), 449–465. https://doi.org/10.2307/2669386.

Natekin, A. & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 21. https://doi.org/10.3389/fnbot.2013.00021.

Neyman, J. & Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 20(1), 1–29. https://doi.org/10.1111/j.2517-6161.1958.tb00272.x.

NVBR (2010). De brandweer over Morgen. *Arnhem: Nederlandse Vereniging voor Brandweer en Rampenbestrijding*.

Nyström, E. J. (1930). Über die prakische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica*, 54, 185–204. https://doi.org/10.1007/BF02547521.

Ogata, Y. & Tanemura, M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Annals of the Institute of Statistical Mathematics*, 33, 315–338. https://doi.org/10.1007/BF02480944.

Ohser, J. & Stoyan, D. (1981). On the second-order and orientation analysis of planar stationary point processes. *Biometrical Journal*, 23(6), 523–533. https://doi.org/10.1002/bimj.4710230602.

Okawa, M., Iwata, T., Kurashima, T., Tanaka, Y., Toda, H. & Ueda, N. (2019). Deep mixture point processes: Spatio-temporal event prediction with rich contextual information. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 373–383. https://doi.org/10.1145/3292500.3330937.

Pawlas, Z. (2011). Estimation of summary characteristics from replicated spatial point processes. *Kybernetika*, 47(6), 880–892. https://www.kybernetika.cz/content/2011/6/880.

Pereira, P., Turkman, K. F., Turkman, M. A. A., Sá, A. & Pereira, J. M. C. (2013). Quantification of annual wildfire risk; A spatio-temporal point process approach. *Statistica*, 73(1), 55–68. https://doi.org/10.6092/issn.1973-2201/3985.

Pimont, F., Fargeon, H., Opitz, T., Ruffault, J., Barbero, R., Martin-StPaul, N., Rigolot, E., RiviÉre, M. & Dupuy, J.-L. (2020). Prediction of regional wildfire activity in the probabilistic Bayesian framework of Firelihood. *Ecological Applications*, 31(5), e02316. https://doi.org/10.1002/eap.2316.

Preisler, H. K., Brillinger, D. R., Burgan, R. E. & Benoit, J. W.(2004). Probability based models for estimation of wildfire risk. *International Journal of Wildland Fire*, 13(2), 133–142. https://doi.org/10.1071/WF02061.

Preston, C.(1975). Spatial birth and death processes. *Advances in Applied Probability*, 7(3), 465–466. https://doi.org/10.1017/S0001867800040726.

Prokešová, M. & Vedel Jensen, E. B. (2013). Asymptotic Palm likelihood theory for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 65, 387–412. https://doi.org/10.1007/s10463-012-0376-7.

Rathbun S. L. & Cressie N. (1994). Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes *Advances in Applied Probability*, 26(1), 122–154. https://doi.org/10.2307/1427583.

Rathbun, S. L., Shiffman, S. & Gwaltney, C. J. (2007). Modelling the effects of partially observed covariates on Poisson process intensity. *Biometrika*, 94(1), 153–165. https://doi.org/10.1093/biomet/asm009.

Rathbun, S. L. (2013). Optimal estimation of Poisson intensity with partially observed covariates. *Biometrika*, 100(1), 277–281. https://doi.org/10.1093/biomet/ass069.

Reitzner, M. & Schulte, M. (2013). Central limit theorems for U-statistics of Poisson point processes. *Annals of Probability*, 41(6), 3879–3909. https://doi.org/10.1214/12-AOP817.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4), 731–792. https://doi.org/10.1111/1467-9868.00095.

Ripley B. D. & Kelly, F. P. (1977). Markov point processes. *Journal of the London Mathematical Society*, 15(2), 188–192. https://doi.org/10.1112/jlms/s2-15.1.188.

Ripley, B. D. (2012). *Statistical Inference for Spatial Processes*, Cambridge University Press. https://doi.org/10.1017/CBO9780511624131.

Rodrigues, M. & De la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, 57, 192–201. https://doi.org/10.1016/j.envsoft.2014.03.003.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Sakr, G. E., Elhajj, I. H., Mitri, G. & Wejinya, U. C. (2010). Artificial intelligence for forest fire prediction. In *2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, 1311–1316. https://doi.org/10.1109/AIM.2010.5695809.

Samo, Y.-L. K. & Roberts, S. (2015). Scalable nonparametric Bayesian inference on point processes with Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, 2227–2236. https://dl.acm.org/doi/10.5555/3045118.3045355.

Satir, O., Berberoglu, S. & Donmez, C. (2016). Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomatics, Natural Hazards and Risk*, 7(5), 1645–1658. https://doi.org/10.1080/19475705.2015.1084541.

Schoenberg, F. P. (2005). Consistent parametric estimation of the intensity of a temporal point process. *Journal of Statistical Planning and Inference*, 128(1), 79–93. https://doi.org/10.1016/j.jspi.2003.09.027.

Schoenberg, F. (2023). Some statistical problems involved in forecasting and estimating the spread of SARS-CoV-2 using Hawkes point processes and SEIR models. *Environmental and Ecological Statistics*, 30, 851–862. https://doi.org/10.1007/s10651-023-00591-6.

Schonlau, M. & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 20(1), 3–29. https://doi.org/10.1177/1536867X20909688.

School, M. L. (2018). A log-Gaussian Cox process for predicting chinmey fires at fire department Twente. *University of Twente, Master Thesis*. https://essay.utwente.nl/76579.

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Avalos, C. & Rue, H. (2014). Spatio-temporal log-Gaussian Cox process for modelling wildfire occurrence: the case of Catalonia, 1994-2008. *Environmental and Ecological Statistics*, 21(3), 531–563. https://doi.org/10.1007/s10651-013-0267-y.

Shchur, P., Türkmen, A. C., Januschowski, T. & Günnemann, S. (2021). Neural temporal point processes: A review. In *Proceedings of the the 30th International Joint Conference on Artificial Intelligence*, 4585–4593. https://doi.org/10.24963/ijcai.2021/623.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 43(3), 310–313. https://doi.org/10.1111/j.2517-6161.1981.tb01676.x.

Silverman, B. W. (1998). *Density Estimation for Statistics and Data Analysis* (1st ed.), Routledge. https://doi.org/10.1201/9781315140919.

Sisson S. A. & Fan Y. (2011). Likelihood-free MCMC. In *Handbook of Markov Chain Monte Carlo*, 12, CRC Press. https://doi.org/10.1201/b10905.

Sørensen, M. (1999). On asymptotics of estimating functions. *Brazilian Journal of Probability and Statistics*, 13(2), 111–136. https://www.jstor.org/stable/43600956.

Stojanova, D., Kobler, A., Ogrinc, P., Ženko, B. & Džeroski, S. (2012). Estimating the risk of fire outbreaks in the natural environment. *Data Mining and Knowledge Discovery*, 24, 411–442. https://doi.org/10.1007/s10618-011-0213-2.

Stoyan, D. & Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science*, 15(1), 61–78. https://doi.org/10.1214/ss/1009212674.

Strauss, D. J. (1975). A model for clustering. *Biometrika*, 63(2), 467–475. https://doi.org/10.1093/biomet/62.2.467.

Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, 25. https://doi.org/10.1186/1471-2105-8-25.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 307. https://doi.org/10.1186/1471-2105-9-307.

Strobl, C. & Zeileis, A. (2008). Danger: High power! – Exploring the statistical properties of a test for random forest variable importance. In *COMPSTAT 2008 – Proceedings in Computational Statistics*, 59-66. https://doi.org/10.5282/ubm/epub.2111.

Strobl, C., Hothorn, T. & Zeileis, A. (2009). Party on! *The R Journal*, 1(2), 14–17. https://doi.org/10.32614/RJ-2009-013.

Swain, A. & Clark, D. (2011). Bayesian estimation of the intensity for independent cluster point processes: An analytic solution. *Procedia Environmental Sciences*, 4, 56–61. https://doi.org/10.1016/j.proenv.2011.07.011.

Takacs, R. (1986). Estimator for the pair-potential of a Gibbsian point process. *Statistics*, 17(3), 429–433. https://doi.org/10.1080/02331888608801956.

Tanaka, U., Ogata, Y. & Stoyan, D. (2008). Parameter estimation and model selection for Neyman-Scott point processes. *Biometrical Journal*, 50(1), 43–57. https://doi.org/10.1002/bimj.200610339.

Tanaka, U. & Ogata, Y. (2014). Identification and estimation of superposed Neyman-Scott spatial cluster processes *Annals of the Institute of Statistical Mathematics*, 66, 687–702. https://doi.org/10.1007/s10463-013-0431-z.

Tang, Z. (2020). A pair correlation function for non-stationary spatial point processes that depends on covariates. *Florida State University, PhD Thesis*. https://purl.lib.fsu.edu/diginole/2020_Summer_Fall_Tang_fsu_0071E_16097.

Thurman, A. L. & Zhu, J. (2014). Variable selection for spatial Poisson point processes via a regularization method. *Statistical Methodology*, 17, 113–125. https://doi.org/10.1016/j.stamet.2013.08.001.

Thurman, A. L., Fu, R., Guan, Y. & Zhu, J. (2015). Regularized estimating equations for model selection of clustered spatial point processes. *Statistica Sinica*, 25(1), 173–188. http://dx.doi.org/10.5705/ss.2013.208w.

Turner, R. (2009). Point patterns of forest fire locations. *Environmental and Ecological Statistics*, 16, 197–223. https://doi.org/10.1007/s10651-007-0085-1.

Unwin, D. I. (1996). GIS, spatial analysis and spatial statistics. *Progress in Human Geography*, 20(4), 540–551. https://doi.org/10.1177/0309132596020004.

Vaart, A. W. van der (1998). *Asymptotic Statistics*, Cambridge University Press. https://doi.org/10.1017/CBO9780511802256.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics With S*, Springer. https://doi.org/10.1007/978-0-387-21706-2.

Ver Hoef, J. M. (2012). Who invented the Delta method? *The American Statistician*, 66(2), 124–127. https://doi.org/10.1080/00031305.2012.687494.

Verdoy, P. J. (2019). Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. Choosing the best mesh for each database. *Communications in Statistics - Simulation and Computation*, 50(10), 2990—3030. https://doi.org/10.1080/03610918.2019.1618473.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics*, 63(1), 252–258. https://doi.org/10.1111/j.1541-0420.2006.00667.x.

Waagepetersen, R. (2008). Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika*, 95(2), 351–363. https://doi.org/10.1093/biomet/asn020.

Waagepetersen, R. & Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3), 685–702. https://doi.org/10.1111/j.1467-9868.2008.00702.x.

Walder, C. J. & Bishop, A. N. (2017). Fast Bayesian intensity estimation for the permanental process. In *Proceedings of the 34th International Conference on Machine Learning*, 3579–3588. https://dl.acm.org/doi/10.5555/3305890.3306051.

Wang, Y., Degleris, A., Williams, A. & Linderman, S. W. (2023). Spatiotemporal clustering with Neyman-Scott processes via connections to Bayesian nonparametric mixture models. *Journal of the American Statistical Association*, 119(547), 2382–2395. https://doi.org/10.1080/01621459.2023.2257896.

Wang, H. & Lin, W. (2023). Nonasymptotic theory for two-layer neural networks: Beyond the bias-variance trade-off. *arXiv:2106.04795*. https://doi.org/10.48550/arXiv.2106.04795.

Wendels, M. (2018). A spatio-temporal point process model for firemen demand in Twente. *University of Twente, Bachelor Thesis*. https://essay.utwente.nl/72216.

Wongvibulsin, S., Wu, K. C. & Zeger, S. L. (2019). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Medical Research Methodology*, 20, 1. https://doi.org/10.1186/s12874-019-0863-0.

Xie, F. & Xu, Y. (2020). Bayesian repulsive Gaussian mixture model. *Journal of the American Statistical Association*, 115(529), 187–203. https://doi.org/10.1080/01621459.2018.1537918.

Xu, G., Waagepetersen, R. & Guan, Y. (2019). Stochastic quasi-likelihood for case-control point pattern data. *Journal of the American Statistical Association*, 114(526), 631–644. https://doi.org/10.1080/01621459.2017.1421543.

Xu, G., Liang, C., Waagepetersen, R. & Guan, Y. (2023). Semiparametric goodness-of-fit test for clustered point processes with a shape-constrained pair correlation function. *Journal of the American Statistical Association*, 118(543), 2072–2087. https://doi.org/10.1080/01621459.2022.2029456.

Xu, H. & Schoenberg, F. P. (2011). Point process modeling of wildfire hazard in Los Angeles county, California. *Annals of Applied Statistics*, 5(2), 684–704. https://doi.org/10.1214/10-AOAS401.

Xu, H., Luo, D. & Zha, H. (2017). Learning Hawkes processes from short doubly-censored event sequences. In *Proceedings of the 34th International Conference on Machine Learning*, 3831–3840. https://dl.acm.org/doi/10.5555/3305890.3306077.

Yang, J., Weisberg, P. J., Dilts, T. E., Loudermilk, E. L., Scheller, R. M., Stanton, A. & Skinner, C. (2015). Predicting wildfire occurrence distribution with spatial point process models and its uncertainty assessment: A case study in the Lake Tahoe Basin, USA. *International Journal of Wildland Fire*, 24(3), 380–390. https://doi.org/10.1071/WF14001.

Yau, C. Y. & Loh, J. M. (2012). A generalization of the Neyman-Scott process. *Statistica Sinica*, 22(4),1717–1736. http://doi.org/10.5705/ss.2010.245.

Yin, L. & Sang, H. (2021). Fused spatial point process intensity estimation with varying coefficients on complex constrained domains. *Spatial Statistics*, 46, 100547. https://doi.org/10.1016/j.spasta.2021.100547.

Yin, F., Jiao, J., Yan, J. & Hu, G. (2022). Bayesian nonparametric learning for point process with spatial homogeneity: a spatial analysis of NBA shot locations. In *Proceedings of the 39th International Conference on Machine Learning*, 25523–25551. https://proceedings.mlr.press/v162/yin22a.html.

Yue, Y. R. & Loh, J. M. (2011). Bayesian semiparametric intensity estimation for inhomogeneous spatial point processes. *Biometrics*, 67(3), 937–946. https://doi.org/10.1111/j.1541-0420.2010.01531.x.

Yue, Y. R. & Loh, J. M. (2015). Variable selection for inhomogeneous spatial point process models. *Canadian Journal of Statistics*, 43(2), 288–305. https://doi.org/10.1002/cjs.11244.

Zhang, J., Cai, B., Zhu, X., Wang, H., Xu G. & Guan Y.(2023). Learning human activity patterns using clustered point processes with active and inactive states. *Journal of Business & Economic Statistics*, 41(2), 388–398. https://doi.org/10.1080/07350015.2021.2025065.

Zhang, Y., Kong, Q. & Zhou, F. (2023). Integration-free training for spatio-temporal multimodal covariate deep kernel point processes. In *Proceedings of the 37th International Conference on Neural Information Processing System*, 25031–25049. https://dl.acm.org/doi/abs/10.5555/3666122.3667210.

Zhu, S. & Xie, Y. (2022). Spatiotemporal-textual point processes for crime linkage detection. *Annals of Applied Statistics*, 16(2), 1151–1170. https://doi.org/10.1214/21-AOAS1538.

Zuo, S., Jiang, H., Li, Z., Zhao, T. & Zha, H. (2020). Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, 11692–11702. https://dl.acm.org/doi/10.5555/3524938.3526022.

Summary

Point pattern data are ubiquitous in both the natural environment and human activities. Spatial and spatio-temporal point processes are widely utilized to model the occurrence of random events across space and time. These models aim to uncover the underlying patterns of such events by analyzing environmental variables that influence their probabilities, and by testing, estimating and simulating spatial or spatio-temporal correlations. Despite significant advances in point process theory and methodology in the last decades, classical approaches still require continuous refinement and extension. This need arises from the distinctive characteristics of point pattern data in practical applications, as well as the increasing amount and complexity of data available in contemporary research. This thesis explores spatial and spatio-temporal modelling in the context of fire risk prediction in the Netherlands. Fire occurrences exhibit unique features that pose specific challenges in point process modelling. Four critical research questions are identified based on real applications of point process methods to the data. To address these challenges, we developed novel statistical and machine learning approaches, proposing targeted solutions that advance the state-of-the-art in the research domain.

Chapter 1 of this thesis employs chimney fire risk prediction as a case study to examine the difficulties of feature selection in point process modelling when faced with a large number of covariates. Traditional approaches often incorporated only a limited number of covariates, neglecting the practical complexities introduced by the numerous potential environmental variables. While some studies addressed variable selection, they typically relied on the estimation of covariate importance through specific statistical metrics, often assuming a parametric framework for the point process of interest in advance. This assumption limits the ability to capture nonlinear relationships and exacerbates the issue of interference among correlated variables. To address these limitations, we combine machine learning techniques with point process modelling to propose a novel hybrid approach. Initially, random forests are used to perform data-driven variable selection, identifying the most important covariates without reliance on parametric assumptions. Subsequently, a parametric point process model is constructed and analyzed based on the selected variables. This method offers several advantages: it avoids the limitations of parametric assumptions during variable selection, ensures interpretability by utilizing classical statistical methods for model inference, and facilitates uncertainty quantification. The proposed approach demonstrates excellent predictive performance in chimney fire risk prediction. Additionally,

this framework is extended to other fire prediction projects, highlighting its generalizability and effectiveness in addressing similar modelling challenges.

Chapter 2 addresses uncertainty analysis in the parameter estimation of intensity functions of point process models, an important component for completely predicting the occurrences of random events. Uncertainty analysis facilitates the construction of confidence intervals for estimated models, enhancing their reliability and interpretability. While existing research has explored uncertainty analysis for point process models, most studies provide asymptotic results based on the increasing-domain framework. Increasing-domain and infill asymptotic regimes represent two principle frameworks for uncertainty analysis for point process models. The former indicates that the accuracy of parameter estimation improves as more events are observed in an increasing window. However, in some cases, point pattern data may exhibit domain-specific characteristics, e.g. fire patterns may differ in urban and rural areas. In such cases, it is more appropriate to adopt the infill asymptotic framework, which fixes the window while observing more and more events across different observation intervals. In this chapter, we systematically study the asymptotic distribution of estimated parameters under the infill regime. For logistic regression estimators, as well as more general estimators derived from unbiased estimating equations, we provide comprehensive asymptotic results applicable to both Poisson and general point processes of interest. In addition, we develop a consistent estimator for the covariance matrix, enabling practical computation of confidence intervals. This theoretical advancement completes the statistical literature on uncertainty quantification in scenarios where domain-specific characteristics require a fixed observation window.

Chapter 3 builds upon the work presented in Chapter 1, extending the proposed methodology for handling a large number of covariates in point process modelling. Although Chapter 1 introduces a data-driven modelling procedure for variable selection on medium-sized point pattern data, it still exhibits several limitations when applied to practical scenarios. The covariates identified as important by machine learning techniques in Chapter 1 are concluded at an aggregation level of areal unit data, which may be different from those concluded at the level of point pattern data. Moreover, complex, potentially non-linear relationships between covariates and the intensity function of a point process bring significant difficulties in constructing appropriate parametric models. Parametric models impose restrictive assumptions that may not hold in real-world point patterns. To overcome these limitations, we propose a nonparametric tree-based method for estimating the intensity functions of point processes, leveraging the flexible function approximation capability of the XGBoost system. To adapt the standard XGBoost framework to point process modelling, we employ a Poisson likelihood-based loss function for model fitting. To improve estimation efficiency for clustered point processes of interest, we analyze the estimation variance of the estimated intensities and design a dynamic weighted loss function in order to align with the boosting mechanism. The proposed method is validated on

both simulated and real-world data sets and demonstrated superior performance compared to existing state-of-the-art approaches. This methodological advancement offers improved accuracy and flexibility in modelling point process intensity functions.

Chapter 4 studies a broad class of hierarchical cluster point processes, known as independent cluster point processes, which are particularly suitable for analyzing point pattern data with hidden cluster centres. These processes comprise two components: a parent point process that generates cluster centers and a mark process associated with each cluster that simulates the observed events. In the literature, simplified versions of these processes, such as Neyman-Scott processes, have been well investigated, where both parent and mark processes are assumed to Poisson point processes. Moment-based approaches have been proposed to estimate model parameters. However, in more general processes whose parent and mark processes are not Poisson, classical estimation methods face difficulties due to the complex probability density function of the model, which involves high-dimensional integrals in high-order summary statistics. An additional challenge arises in estimating the distribution of cluster centres, also referred to as state estimation in point process modelling. This task is particularly crucial in practical applications like arson fire risk prediction, where identifying cluster centres provides valuable prior information for the police and firefighters to guide prevention and rescue operations. Frequentist methods, however, are inherently limited in this context, as they cannot construct posterior density surfaces for cluster centres. In this chapter, we propose a two-step Bayesian inference framework. In the first step, we employ Markov chain Monte Carlo sampling to estimate the cluster centres, conditional on certain given model parameters. In the second step, these samples were utilized within a Monte Carlo expectation-maximization algorithm to estimate the true model parameters. By iteratively alternating these two steps, the independent cluster point process model can be effectively resolved. Simulation studies demonstrate the effectiveness of the proposed method, and its application to real-world data is currently underway.

In conclusion, this thesis brings novel contributions to the development of spatial and spatio-temporal point process modelling. It establishes methodologies and theories to address significant challenges in the research domain. In future work, we aim to refine these methodologies and theories to better accommodate specific characteristics and complexities of real-world point pattern data.

Samenvatting

Puntpatronen zijn alomtegenwoordig, zowel in de natuur als ter beschrijving van menselijke activiteiten. Om dit type gegevens te modelleren, gebruikt men in de wiskunde puntprocessen in het platte vlak of in ruimte-tijd. In het model kan de invloed van omgevingsvariabelen op het voorkomen van een punt worden meegenomen, alsmede de onderlinge afhankelijkheid van punten. Vervolgens kan men dan een hypothese toetsen, modelparameters schatten of puntpatronen simuleren.

Ondanks dat er in de afgelopen decennia aanzienlijke vooruitgang is geboekt in de theorie van puntprocessen, vereisen klassieke benaderingen voortdurende verfijning en uitbreiding. Deze noodzaak vloeit voort uit de specifieke kenmerken van een gegeven toepassing, evenals uit de toenemende complexiteit van de gegevens die voor hedendaags onderzoek beschikbaar zijn. Dit proefschrift verkent modellering in ruimte en tijd binnen de context van brandrisicovoorspelling in Nederland. Wij formuleren vier onderzoeksvragen gemotiveerd door de praktijk en ontwikkelen nieuwe statistische en machine learning-technieken teneinde gerichte oplossingen te bieden.

Hoofdstuk 1 bestudeert de casus van het voorspellen van het risico op schoorsteenbrand in Twente aan de hand van een groot aantal omgevingsvariabelen. De onderzoeksvraag is om te bepalen welke variabelen van belang zijn, en welke kunnen worden genegeerd. Traditionele benaderingen gaan uit van een beperkt aantal variabelen binnen een parametrisch statistisch kader. Voor onze casus met vele, onderling gecorreleerde variabelen combineren wij het statistisch kader met niet-parametrische technieken uit de machine learning. Wij gebruiken willekeurige beslissingsbossen om de belangrijkste variabelen te vinden. Vervolgens wordt een parametrisch model opgesteld en geanalyseerd op basis van de geselecteerde variabelen. Deze aanpak biedt belangrijke voordelen: hij vermijdt de beperkingen van parametrische modelaannames bij het selecteren van variabelen, waarborgt de interpreteerbaarheid door gebruik te maken van klassieke statistiek en kwantificeert de onzekerheid. In de genoemde casus werden uitstekende resultaten behaald. Bovendien kan de geschetste methode worden uitgebreid naar andere veiligheidsregio's.

Hoofdstuk 2 is gewijd aan een analyse van de onzekerheid bij het schatten van de parameters van de intensiteitsfunctie van een puntproces. Bestaand onderzoek richt zich vooral op asymptotische resultaten voor het regime waarin het waarnemingsvenster groter en groter wordt. Soms echter vertonen puntpatronen domeinspecifieke kenmerken. De patronen van brandincidenten in stedelijke en

landelijke gebieden, bijvoorbeeld, zijn verschillend. In dergelijke gevallen is het beter om een asymptotisch regime te hanteren waarin meer en meer gebeurtenissen worden waargenomen binnen hetzelfde venster. In dit hoofdstuk bestuderen we de asymptotische verdeling van logistische regressieschatters (en andere M-schatters) onder laatstgenoemd regime. Wij presenteren resultaten voor zowel Poissonprocessen als puntprocesmodellen in het algemeen. Daarnaast is een asymptotisch rake schatter voor de covariantiematrix ontwikkeld.

Hoofdstuk 3 bouwt voort op het werk van Hoofdstuk 1. In dat hoofdstuk was het mogelijk om met de geselecteerde variabelen een interpreteerbaar parametrisch statistisch model op te stellen. In het algemeen echter kan de relatie tussen omgevingsvariabelen en de intensiteitsfunctie van een puntproces complex zijn. Voor dergelijke gevallen stellen wij een niet-parametrische, op een beslissingsbos gebaseerde methode voor, waarbij wij gebruikmaken van het flexibele XGBoost-kader voor het benaderen van functies. Om deze methode aan te passen aan onze context, gebruiken we een verliesfunctie gebaseerd op de Poisson waarschijnlijkheidsfunctie. Voor processen die gekenmerkt worden door ophopingen van punten kan de verliesfunctie dynamisch worden gewogen. De nieuwe methode is vergeleken met de huidige stand der techniek en superieur bevonden.

Hoofdstuk 4 betreft de klasse der onafhankelijke clusterprocessen, die bijzonder geschikt is voor het analyseren van puntpatronen met verborgen informatie. Onze motivatie bestaat uit het opsporen van pyromanen aan de hand van gerapporteerde aangestoken autobranden. Voor specifieke modellen gebaseerd op een Poissonaanname voor zowel het verborgen als waargenomen puntproces bestaat een uitgebreide literatuur. In het algemeen is de kansverdeling van het proces echter niet expliciet te beschrijven. Ook het schatten van de verdeling van het verborgen proces is uitdagend, omdat frequentistische methoden hiervoor niet geschikt zijn. In dit hoofdstuk ontwikkelen wij daartoe een twee-staps Bayesiaans statistisch kader. In de eerste stap gebruiken we Monte Carlo methoden om de kansverdeling van het verborgen proces te schatten op basis van de waarnemingen. In de tweede stap gebruiken we monsters uit deze kansverdeling binnen een Monte Carlo EM-algoritme om de modelparameters te schatten. Een simulatiestudie bewijst de effectiviteit van de voorgestelde methode; toepassing op autobrandgegevens is in uitvoering.

Concluderend kunnen wij stellen dat dit proefschrift bijdragen levert aan het modelleren en de analyse van puntprocessen met specifieke toepassingen binnen het onderzoek naar brandincidenten.

Author's Information

Changqing Lu was born on November 2, 1997, in Taizhou, Jiangsu, China. In 2015, he began his undergraduate studies in Automation, with a focus on Machine Learning, in the honor's programme at Northwestern Polytechnical University in Xi'an, Shaanxi, China. During his bachelor, he was awarded National Scholarship multiple times by Ministry of Education in China and was selected for the joint education programme between Northwestern Polytechnical University and the University of Twente. Through this programme, he earned a master degree in Computer Science, with a specialization in Data Science and Technology. He also won a scholarship for outstanding students in this programme.

In September 2020, Changqing began his PhD with the Stochastic Operations Research group at the University of Twente, under the supervision of prof. dr. Marie-Colette van Lieshout. His research focuses on statistical and machine learning methods for spatial and spatio-temporal point process modelling. During his PhD, he conducted a research visit to prof. dr. Yongtao Guan at the Chinese University of Hong Kong and was invited to present at several prestigious international conferences. Recently, he obtained the 'Alain Bensoussan' fellowship, funded by the European Research Consortium for Informatics and Mathematics, and will soon start a postdoctoral position at the National Research Institute for Mathematics and Computer Science (CWI) in the Netherlands.